

THE
BRITISH JOURNAL
OF
STATISTICAL
PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ALAN STUART
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN	E. A. PEEL
M. S. BARTLETT	L. S. PENROSE
W. G. EMMETT	J. FRASER ROBERTS
M. G. KENDALL	A. RODGER
D. F. LAWLEY	W. STEPHENSON
E. S. PEARSON	P. E. VERNON

Managing Sub-editor J. W. WHITFIELD

Printed and Published by
TAYLOR & FRANCIS LTD.

RED LION COURT, FLEET STREET, LONDON, E.C.4

Price 20s. per part (U.S.A. \$3.50)

Subscription 30s. 6d. per volume (U.S.A. \$4.50)



222
9/8

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Statistical Psychology is issued by the British Psychological Society. The subscription price per volume is 30s. net for Members of the Society and 30s. 6d. (post free) for non-members. The subscription price in the U.S.A. is \$4.50 (post free). Members of the Society should send their subscriptions to THE SECRETARY, British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1: non-members to Messrs. TAYLOR & FRANCIS, LTD., 18 Red Lion Court, Fleet Street, London, E.C.4. Until further notice the Journal will be issued in two parts each year, in May and November.

Papers for publication should be sent to Sir CYRIL BURT, 9 Elsworth Road, London, N.W.3. Authors will receive 25 copies of their articles free; extra copies may be purchased provided the order is given when the proofs are returned. Contributors should indicate whether they wish their reprints to be bound in covers. The printers will state current prices when sending out proofs.

Books for review should be sent to the Editor, advertisements to the Secretary, British Psychological Society.

The Society also issues quarterly *The British Journal of Psychology* and *The British Journal of Medical Psychology*: subscriptions, 60s. per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society publishes, jointly with the Association of Teachers in Colleges and Departments of Education, *The British Journal of Educational Psychology*: for this the subscription, £1 per annum, should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2. Members of the Society receive the foregoing journals on special terms; enquiries should be addressed to the Secretary, British Psychological Society.

Papers for publication in *The British Journal of Psychology* should be sent to Professor James Drever, Department of Psychology, the University, Edinburgh; those for publication in *The British Journal of Medical Psychology* to Dr. Joseph Sandler, 96 Portland Place, London, W.1; those for publication in *The British Journal of Educational Psychology* to Professor P. E. Vernon, London University Institute of Education, Malet Street, London, W.C.1.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in previous publications of this Journal. Each article should be headed either with a series of numbered headings in italic, corresponding with the cross-headings of the several sections, or with a brief abstract: (the latter procedure is preferred). Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. (particularly in the case of tables and references). In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. II, pp. 21-73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, XL, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Manuscripts involving many tables and numerous equations can as a rule only be accepted if the authors are willing to pay for the additional cost.

In future authors will receive two complete slip-proofs of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

A STUDY OF THE PERFORMANCE OF 2,000 CHILDREN ON FOUR VOCABULARY TESTS

II. NORMS, WITH SOME OBSERVATIONS ON THE RELATIVE VARIABILITY OF BOYS AND GIRLS

By M. I. DUNSDON and J. A. FRASER ROBERTS

Burden Mental Research Department, Stoke Park Hospital, Bristol

I. Introduction. II. Increment of Mean Score with Age. III. Increment of Variance of Score with Age. IV. Norms. V. Efficiency of Adjusted Scores. VI. Frequency Distribution of Vocabulary Score. VII. Further Examination of Variability of Score in Boys and Girls. VIII. Summary.

I. INTRODUCTION

Reasons have been given (Dunsdon and Roberts [1]) for choosing vocabulary scales as quickly-given individual tests, for the purpose, not of assessing single children, but of comparing large groups. For the comparison of groups vocabulary tests have manifest advantages in respect of reliability, validity, relative absence of practice effect, and above all, the amount of information obtained per unit of testing time. The four oral definition vocabularies used here are those included in the Terman-Merrill (1937 Revision) [8] Form L, and Wechsler's Intelligence Scale for Children [9], together with Raven's Mill Hill Vocabulary Tests A and B [4]. In the form for oral definition the Mill Hill vocabularies were designed by Raven to be used in combination. Accordingly in this paper results and norms are given in terms of the combined scores for M.H.A. and M.H.B. Norms are therefore presented, by sexes, for

1. Terman-Merrill Intelligence Scale, Form L (T.M.L.);
2. Wechsler Intelligence Scale for Children (W.I.S.C.);
3. Mill Hill A and Mill Hill B Vocabulary Scales combined (M.H.A. and M.H.B.);
4. All four vocabularies combined.

The reason for calculating norms is to derive scores which are independent of age and sex. The requirements for assessing individuals on the one hand and comparing large groups on the other are very different. For the latter a much higher standard of accuracy is needed. Relatively small systematic sources of bias may matter little in an individual measurement, but are important when the problem is (to quote a concrete example) whether twins are or are not slightly lower in average performance than their sibs, who are necessarily of different age. The point is well illustrated in a recent valuable review article by Yates [10], dealing with the use of vocabulary in the measurement of intellectual deterioration.

S.P.

Bureau of Ednl. & Psyl. Research
(S. C. E. R. I.)

Date

T 943

Bureau Ednl. & Psyl. Research
MILL HILL COLLEGE

A

He mentions the higher average score of boys, quoting our study and others; and then adds: "The differences were so slight as to be unimportant". Actually, on taking the four vocabularies combined, at 6 years boys have a mean score of 26 words and girls 23; at 14 years boys score 76 and girls 71. In terms of 'mental ages' these are differences of about 5 months in young children and nearly a year in children approaching 15. Such differences may be unimportant in relation to the special problem Yates discusses; but they are of the order of magnitude of some of the differences between large groups of subjects that we had in mind when undertaking this work. The same considerations would apply to individual results if vocabulary tests were being used as part of a selection procedure for children too young to be tested with group scales. Norms must in fact be calculated separately for boys and girls.

Similar considerations apply to standardization for differences in age. Rough norms are not enough. Only if it can be shown that the adjustments do in fact yield scores which are independent, or almost independent, of differences in age and sex, can groups be compared which differ in age and sex composition. Matching becomes necessary, the very complication that norms are designed to avoid. We have tried to adopt a procedure which uses as much as possible of the information provided by the sample; and we feel that, in view of the purpose of the work, this needs no apology. In any case, more accurate norms, though much more troublesome to compute, are easy enough to use.

The first step in the procedure which we have adopted is to measure by how much the number of words correctly defined differs from the number expected for that sex and that age, the expected number being calculated from the fitted growth curves. A second step is then required. Variability also increases with advancing age; hence it is necessary to adjust the deviations from the expected scores so that these are made equivalent for children of different ages and incidentally for children of different sex. This can be done by multiplying the deviation at any observed age by the ratio of the standard deviation at that age to some fixed standard deviation. The method is perhaps a little cumbersome; but it has worked well with previous material, and, as will be shown, it also works well in the present instance. Nor does it seem likely that any rational transformation of the unit of measurement could make possible a simpler and more direct approach.

A sex difference in the variability of scores for intelligence tests has often been observed, and raises questions which have frequently been discussed, though certainly never settled. Thanks to its wide range of age, the present sample provides useful material for a further examination of this problem. The results will be presented in Section VII.

II. INCREMENT OF MEAN SCORE WITH AGE

In our previous paper [2] linear equations for increment of score with advancing C.A. were given by sexes for each vocabulary separately and for the four combined. The fit was good, without significant remainder, except for

W.I.S.C. (which in both sexes showed a decreasing rate of increment) and for Mill Hill A (which showed the same phenomenon, though, curiously enough, in girls only). Quadratics, however, gave an excellent fit in all three instances. Although the other remainders were not significant, nevertheless with M.H.A. for boys, M.H.B. for girls and with all four vocabularies combined for girls, quadratic equations improved the fit, the variance ratio for the addition of the quadratic term being significant in all three instances. Where a quadratic is used for one sex, it is desirable to use a second-order curve for the other sex as well, even though the improvement may be trivial; otherwise the sex difference at the extremes of the range of age may be exaggerated. For this reason, and for the sake of uniformity, quadratics have been used throughout. The importance of securing a good fit for the growth curves of the means, and the substantial improvement effected in some instances by the use of quadratics, is well brought out in the final testing of the norms by means of contingency tables, as described in Section V.

TABLE I. W.I.S.C. VOCABULARY
Quadratic Curves for Increment of Score. Analysis of Variance

Variation of Vocabulary Score	D. of F.	Sum of Squares (words) ²	Mean Square (words) ²	Variance Ratio	Significance
Boys					
Linear regression	1	15,858.3	15,858.3		
Addition of quadratic term	1	90.6	90.6	7.95	++
Remainder between arrays	17	288.0	16.94	1.50	—
Within arrays	960	10,839.3	11.29		
Total	979	27,076.2			
Girls					
Linear regression	1	14,086.3	14,086.3		
Addition of quadratic term	1	206.6	206.6	23.34	++
Remainder between arrays	17	101.5	5.97	0.67	—
Within arrays	947	8,427.0	8.90		
Total	966	22,821.4			

The most important discrepancy to be corrected occurred with W.I.S.C.; the result of fitting quadratics is shown in Table I. Although M.H.A. is not used alone, mention may be made of the disappearance of the previously highly significant remainder shown by the girls. The variance ratio for the difference between the linear and curvilinear regressions, with degrees of freedom 1 and 964, is 18.63, and the variance ratio for the remainder, with degrees of freedom 17 and 947, is 1.13, as compared with the previous high figure of 2.10. The fitting of the growth curves for M.H.A. and M.H.B. combined is shown in Table II.

TABLE II. M.H.A. + M.H.B. VOCABULARIES
Quadratic Curves for Increment of Score. Analysis of Variance

Variation of Vocabulary Score	D. of F.	Sum of Squares (words) ²	Mean Square (words) ²	Variance Ratio	Significance
Boys					
Linear regression	1	77,810.3			
Addition of quadratic term	1	135.2	135.2	3.23	—
Remainder between arrays	17	906.0	53.29	1.28	—
Within arrays	960	39,996.3	41.66		
Total	979	118,847.8			
GIRLS					
Linear regression	1	76,542.3			
Addition of quadratic term	1	394.2	394.2	12.10	++
Remainder between arrays	17	531.5	31.27	0.96	—
Within arrays	947	30,875.8	32.60		
Total	966	108,343.8			

The improvement with the four vocabularies combined is rather small. But the quadratic term is significant with the girls; and the variance ratio for the remainder is reduced from 1.20 to 0.86.

The regression equations used for calculating expected scores are as follows :

T.M.L.	Boys: $y = -0.95 + 0.0855267x + 0.000216918x^2$, Girls: $y = -1.82 + 0.0976822x + 0.000124580x^2$,
W.I.S.C.	Boys: $y = -1.62 + 0.189335x - 0.000291713x^2$, Girls: $y = -3.69 + 0.215646x - 0.000439474x^2$,
M.H.A. + B	Boys: $y = -14.18 + 0.350433x - 0.000356116x^2$, Girls: $y = -18.01 + 0.404122x - 0.000607071x^2$,
4 Combined	Boys: $y = -16.76 + 0.625292x - 0.000430884x^2$, Girls: $y = -23.52 + 0.717450x - 0.000921966x^2$,

where x = C.A. in completed months and y = score in words.

III. INCREMENT OF VARIANCE OF SCORE WITH AGE

The expected scores by age and sex having been obtained, the next step is to calculate multipliers for making deviations from expectation comparable over the whole range of age, five to fifteen years, and for children of opposite sex. Each multiplier is the ratio of the standard deviation at observed C.A. to some fixed standard deviation. In order to estimate the standard deviations at observed C.A. with as much accuracy as possible curves are fitted for the regression on C.A. of the variance of scores. This procedure was originally used by one of us in connection with norms for the Advanced Otis Scale (Roberts, Norman and Griffiths [5]), and has since been explained more fully (Roberts and Mellone [6]).

Roberts [7]). It is a device that can be applied to any measurements when the range of age is great enough to yield a considerable increase in variance, and when no simple transformation of the units of measurement seems likely to provide a better alternative. For example, it has worked well for measurements of arterial blood pressure (Hamilton, Pickering, Roberts and Sowry [3]).

It was found that cubics are required. In four of the eight comparisons the addition of the cubic term effected significant improvement. What is more important is that straight lines or quadratics give expected variances for the youngest children which are far too low or even negative; this is shown clearly by Fig. 1.

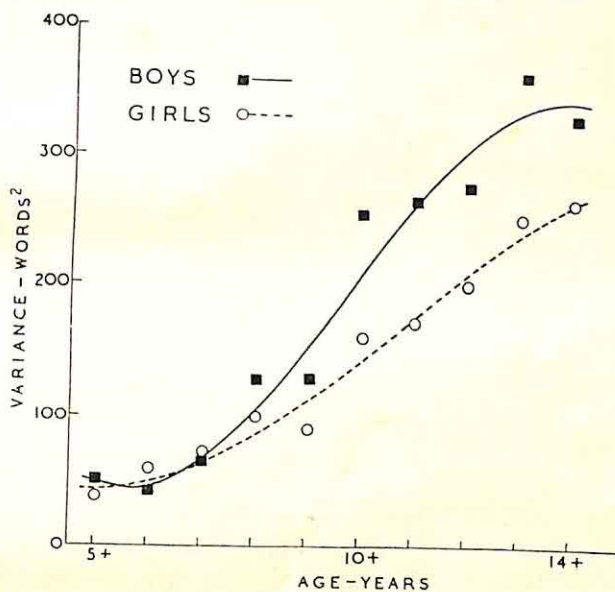


FIG. 1. Four vocabularies combined.
Increment of variance of score with age. Means of year age groups and fitted curves.

TABLE III. REGRESSION OF VARIANCE OF SCORE ON AGE
Significance of Remainders after Fitting Cubics

Vocabulary	BOYS		GIRLS	
	Variance Ratio	Significance	Variance Ratio	Significance
T.M.L.	1.72	+	2.44	++
W.I.S.C.	2.02	++	1.36	-
M.H.A.+M.H.B.	1.94	+	0.61	-
4 combined	1.84	+	0.50	-

Degrees of freedom : 16 and ∞ throughout.

A summary of the results is given in Table III. With the girls the fit is excellent, except for T.M.L. With the boys the fit is not so good, all the remainders being significant, though not very large. It should be mentioned, however, that the final adjusted scores are a good deal less sensitive to changes in expected variances than they are to changes in expected means. Fig. 1, which shows the observed means of arrays and the fitted curves for all four vocabularies combined, brings out the closeness of the fit with the girls and some irregularity with the older boys. These features are also true of the individual vocabularies taken separately.

TABLE IV. FOUR VOCABULARIES COMBINED
Increment of Variance of Score with C.A. Analysis of Variance

Variation of Variance due to	D. of F.	Sum of Squares (words) ⁴	Mean Square (words) ⁴	Theoretical Variance (D. of F. = ∞)	Variance Ratio	Signi- ficance
Boys						
Linear Regression	1	11,508,784	11,508,784	58,342	197.26	++
Addition of quadratic term	1	40,055	40,055	93,885	0.43	—
Addition of cubic term	1	388,507	388,507	94,348	4.12	+
Remainder	16	2,832,823	177,051	96,023	1.84	+
Total	19	14,770,169				
GIRLS						
Linear Regression	1	5,272,883	5,272,883	33,194	158.85	++
Addition of quadratic term	1	71,523	71,523	50,037	1.43	—
Addition of cubic term	1	103,369	103,369	50,537	2.05	—
Remainder	16	408,560	25,535	50,695	0.50	—
Total	19	5,856,335				

Table IV gives, as an example, the analysis of variance for the 4 vocabularies combined. The unit of grouping for C.A. is 6 months. The other analyses were fairly comparable.

The regression equations are as follows:

T.M.L. Boys: $y = 100.88 - 3.02508x + 0.0286268x^2 - 0.0000778981x^3$,
 Girls: $y = 95.31 - 2.68802x + 0.0239035x^2 - 0.0000612672x^3$,
 W.I.S.C. Boys: $y = 60.88 - 1.73925x + 0.0166865x^2 - 0.0000464019x^3$,
 Girls: $y = 15.23 - 0.358567x + 0.00380813x^2 - 0.0000102841x^3$,
 M.H.A. + B Boys: $y = 108.20 - 3.28393x + 0.0335929x^2 - 0.0000890677x^3$,
 Girls: $y = 2.85 - 0.126640x + 0.00404109x^2 - 0.00000704577x^3$,
 4 Combined Boys: $y = 751.17 - 22.7530x + 0.223352x^2 - 0.000609357x^3$,
 Girls: $y = 252.34 - 7.28021x + 0.0745683x^2 - 0.000185767x^3$,
 where x = age in completed months and y is variance of score in (words)².

IV. NORMS

Each of the tables printed in the Appendix is divided into two parts. The first gives for each vocabulary, or group of vocabularies, C.A. in years and months, with the corresponding number of words; from these are read off the number of words by which the observed score exceeds or falls short of expectation. The second gives the multipliers which apply to successive 6-month age-groups. The observed deviation, plus or minus, is multiplied by the appropriate multiplier; and the final adjusted scores are thus made comparable for children of different ages and opposite sex.

The choice of the fixed standard deviations to which adjustment is made is explained in the notes appended to the tables of norms. All have a direct relation to the deviation from expectation in number of words as at some particular age within the range of the sample. We ourselves are using, with one exception, the plus or minus scores just as they stand. They have, of course, a mean of 0 for our sample. With the 4 vocabularies combined, however, a standard deviation of 15 is within the age range; so in this instance we add 100, thereby securing a frequency distribution with the conventional mean of 100 and standard deviation of 15.

One point in the calculation of the multipliers needs explanation. Some of the cubics are rather sharply inflected near the lower end of the range, which means that, if multipliers were given for each 6-month age-group, they would first of all increase rather rapidly from the group 5.0-5.5 before decreasing again. These tails are clearly rather unstable; so it seemed best to average the variances for successive groups, starting with the youngest, until the next older group showed a decrease from those younger groups so far combined. The six-month age-groups for multipliers may be thought rather coarse. With younger children the values are changing rapidly; with the 4 vocabularies combined the largest difference is 0.21, as the multiplier is 1.94 for boys of 7.0-7.5, and 1.73 for boys of 7.6-7.11. When the observed C.A. lies towards the end of a 6-month group, interpolation can be used for greater accuracy, though it seems hardly necessary. In the whole of the sample of 1,947 children, there are only 4 instances where interpolation gives a difference of as much as 2 points of score. To give an example, a girl aged 7.11 defined 66 words, 30 words above expectation; the application of the multiplier gives her an adjusted score of +54; and this is merely reduced to +52 on interpolation.

V. EFFICIENCY OF ADJUSTED SCORES

It is desirable to see whether the adjusted scores are indeed achieving their purpose, that is, whether in fact they are independent of age and sex. Average allowance has been made for mean and variance; but there may be non-normal variations, for example skewness, which are different at different ages or in the two sexes. Accordingly, adjusted scores have been worked out for the 1,947 children making up the sample. Taking the sex comparisons first, the results are shown in Table V; no difference is significant; so that the adjusted scores are comparable for boys and girls.

TABLE V. ADJUSTED SCORES. COMPARISON OF SEXES

Vocabulary	Mean		S.D.	
	Boys	Girls	Boys	Girls
T.M.L.	+0.030	0.000	3.533	3.449
W.I.S.C.	-0.011	+0.021	3.002	3.007
M.H.A. + M.H.B.	-0.045	-0.051	5.803	5.716
4 Vocabularies	+0.006	-0.148	15.034	14.920

All differences non-significant

Independence of C.A. can be tested by forming contingency tables. In view of the results of Table V the scores of boys and girls can be added. The units of grouping are two years of age and 2 words, except for the 4 vocabularies combined, in which 5-word grouping is used. The results are shown in Table VI.

TABLE VI. TESTS OF INDEPENDENCE OF ADJUSTED SCORE AND C.A. : SUMMARY

Vocabulary	D. of F.	Chi Squared	P
T.M.L.	28	78.77	<0.00001
W.I.S.C.	28	44.23	0.03
M.H.A. + M.H.B.	52	54.23	0.39
4 Vocabularies Combined	52	52.13	0.47

It will be seen that M.H.A. + M.H.B. and the 4 vocabularies combined show almost perfect independence of C.A. and score. With W.I.S.C. chi-squared is significant, but not highly so. With T.M.L., on the other hand, there is a large and highly significant departure from independence. If the calculation is repeated, however, leaving out the youngest 2-year group, chi-squared falls to a value below the 5 per cent level of significance. The departure from independence, therefore, does not seem serious, as no one would want to use in isolation a single vocabulary of this kind for very young children.

As might be expected, the substitution of quadratic for linear growth curves for the means makes practically no difference when the improvement is small. Contingency tables based on the linear equations for T.M.L. and for the 4 vocabularies combined give values for chi-squared which are closely similar to those of Table VI. On the other hand, with M.H.A. and M.H.B. there was an appreciable falling off in rate of increment, which leads to a sharp increase in chi-squared from the 54.23 shown in Table VI to 82.24 if linear curves are used. No doubt W.I.S.C. would show the same effect; but we have not made this calculation. This illustrates the importance of securing a good fit and the inefficiency that would result from using approximate straight lines when a growth curve is appreciably curvilinear.

VI. FREQUENCY DISTRIBUTION OF VOCABULARY SCORE

With any intelligence scale designed to grade the whole population it is desirable that the scores should have an approximately normal frequency distribution. The adjusted scores for the 4 vocabularies combined give a single frequency distribution for the whole sample; they are no more than deviations from the regression line, further adjusted to equalize for differing variances at different ages and for different sexes. Departures from normality, such as skewness, will not be affected. Smoothed frequencies, with single words of score, as described in the preceding section must be used; and, although this may somewhat affect the degrees of freedom, the comparison should be broadly admissible.

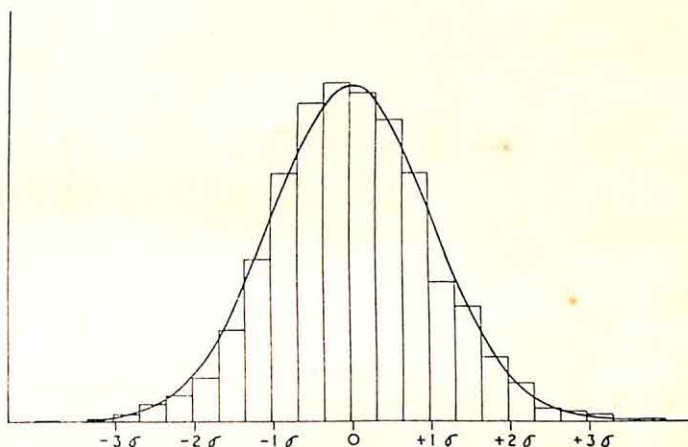


FIG. 2. Four vocabularies combined.

Frequency distribution by 5 point groups of age, and sex adjusted scores, with fitted normal curve.

Fig. 2 shows the observed frequencies in 5-word groups, together with the fitted normal curve. Testing for normality by means of chi-squared, the central 16 5-word groups can be treated separately; the tails below -40 words and above $+39$ words, very nearly $\pm 2\frac{2}{3}\sigma$, must be lumped together to secure sufficient numbers. There are thus 18 arrays and, subject to the reservation made above about smoothing, 15 degrees of freedom. Chi-squared is 28.53, corresponding to a p of very nearly 0.02. The frequency distribution is thus approximately normal, and it therefore appears that scores on well constructed oral definition tests are as normally distributed as can reasonably be expected. The biggest contribution to chi-squared, 8.19, is due to an excess of scores above $+39$. The observed number is 16.75 against an expectation of 8.438. This is a feature of many intelligence tests; it is difficult to design a scale which does not allow a few very bright children to secure unduly high scores. The group $+15$ to $+19$

shows a deficiency; but the general feature that stands out in Fig. 2 is a slight tendency for the frequency curve to be flat-topped at about and rather below the mean. This phenomenon may be real, as it not infrequently appears with scales of different kinds.

VII. FURTHER EXAMINATION OF VARIABILITY OF SCORE IN BOYS AND GIRLS

The question of the relative variabilities of boys and girls in intelligence test scores has attracted interest over many years. The present sample has a wide range of age; and so is useful for examining the problem. The changing pattern as age increases is shown in Table VII. The means and standard

TABLE VII. VARIABILITY OF SCORE

Age (Years)	Mean		S.D.		C.O.V. %		Ratio. Mean	Boys : Girls	
	Boys	Girls	Boys	Girls	Boys	Girls		S.D.	C.O.V.
5+	22.7	20.5	7.3	6.1	32.0	29.7	1.10	1.19	1.08
6+	29.5	27.0	6.5	7.7	21.9	28.6	1.09	0.84	0.76
7+	34.9	31.6	8.2	8.6	23.5	27.2	1.10	0.95	0.86
8+	42.8	39.7	11.4	10.1	26.6	25.4	1.08	1.13	1.05
9+	48.0	45.5	11.4	9.6	23.8	21.0	1.05	1.20	1.13
10+	57.0	53.5	15.9	12.7	27.9	23.8	1.07	1.25	1.17
11+	58.6	57.1	16.2	13.2	27.7	23.1	1.03	1.23	1.20
12+	67.7	64.2	16.6	14.2	24.5	22.1	1.06	1.17	1.11
13+	73.8	66.5	19.1	15.8	25.8	23.8	1.11	1.20	1.08
14+	77.7	73.4	18.2	16.2	23.4	22.1	1.06	1.12	1.06

All figures based on raw scores.

deviations are based on the simple raw scores by year age-groups for all four vocabularies combined. As was shown in the previous paper, boys have a higher mean score at all ages, though, as the eighth column shows, the relative excess tends to decline. With 6 and 7-year-old children the standard deviation is higher in girls, but thereafter is considerably higher in boys. The coefficient of variation gives the clearest picture. The ratios shown in the last column reveal a clear trend. In the youngest age-group, 5.0-5.11, boys have the higher C.O.V.; but it would be unsafe to deduce too much from tests given to such young children. In the seventh year girls are much more variable. At about 8 or 9 years equality is reached; and thereafter the variability of boys relative to that of girls increases steadily to about 11 or 12 years; it then declines just as steadily to age 15.

The ratios of the C.O.V.'s are shown in Fig. 3, where the figures of the last column of Table VII have been plotted. A continuous line is shown which represents the ratio of the C.O.V.'s when calculated from the curves fitted for increment with age of mean and variance. The trends are smoothed; but in fact little is added to the clear evidence of the figures based on the simple raw scores.

It is tempting to conclude that, in relative variability, vocabulary tests may be typical of intelligence scales in general. If so, the greater variability of boys may be a passing phase in development, absent in the young child, reaching a maximum at 11 or 12 years, and perhaps (though this is an extrapolation) disappearing by the time adult life is reached. The phenomenon of greater male variability is particularly striking, because most of the evidence comes from group tests administered at about the age when the difference is maximal. All this is somewhat speculative; but we hope that our data make some contribution, and that others will provide further evidence based on other intelligence scales. The problem is important. If indeed the very real greater male variability is a passing phase, then allowance is rightly made by scoring boys and girls separately in scholarship examinations, and basing grammar school places, for example, on an equal proportion of each sex, and not on a dividing line of score applied to both boys and girls.

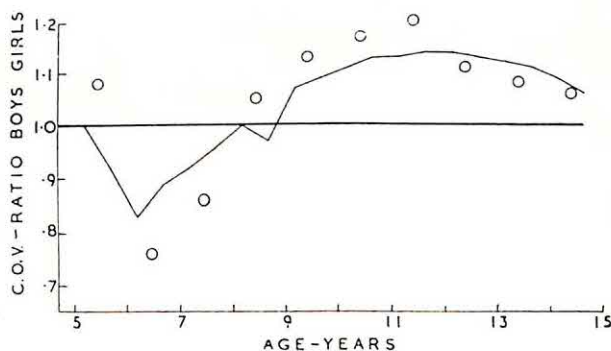


FIG. 3. Four vocabularies combined.

Ratio of coefficients of variation (boys/girls). Circles based on raw scores by year age groups; continuous line based on fitted curves for increment of mean and variance with age.

VIII. SUMMARY

1. A 3 per cent sample of the school children of the City and County of Bristol was selected by visiting all the schools, as well as a number outside, and testing children between the ages of 5.0 to 14.11 years whose homes were within the city and whose birthdays fell on the first day of any calendar month. Four vocabulary scales were used: that from the Terman-Merrill Scale (Form L), Mill Hill Vocabularies A and B, and that from the Wechsler Intelligence Scale for Children.

2. The present paper gives norms for the vocabularies from the Terman-Merrill and Wechsler Intelligence Scales, the oral definitions form of Mill Hill Vocabulary Scales A and B, and for all four vocabularies combined. In view of the sex difference, norms are given separately for boys and girls.

3. Curves have been fitted for the increment of score with advancing age; and tables are given from which deviations from expectation can be read off.

4. Further curves have been fitted for the increment of variance with advancing age. Deviations from expectation can be made comparable for children of different ages by multiplying by the ratio of the standard deviation at observed age to a fixed standard deviation. Tables of multipliers are given for making this further adjustment, which also makes scores comparable for children of opposite sex.

5. It is shown that for this sample the adjusted scores are, with small exceptions, independent of age at test, and that they equalize for children of opposite sex.

6. The frequency distribution is substantially normal.

7. In terms of the coefficient of variation girls were more variable than boys in the younger age-groups. At about 8 or 9 years equality is reached; and thereafter boys become progressively more variable than girls up to the age of 11 or 12. Beyond this age the difference between the sexes is progressively reduced.

ACKNOWLEDGMENTS

We are most grateful to Professor Lewis M. Terman and to the Houghton Mifflin Company for permission to use the Vocabulary test from the Terman-Merrill Intelligence Scale [8]; to Mr. J. C. Raven [4] for permission to use the Oral Definitions Form of his Mill Hill Vocabulary Scale; and to Dr. David Wechsler [9] for permission to use the Vocabulary from his Intelligence Scale for Children. Special thanks are due to the Bristol Education Committee and to the Chief Education Officer, Mr. G. H. Sylvester, who gave permission for work to be carried out in their schools, and in particular to the Headmasters and Headmistresses both of the Local Education Authority's schools and of independent schools, for their willing and active co-operation.

APPENDIX

Notes to Table VIII

1. To obtain scores adjusted for differences in C.A. and sex, read off expected number of words for observed C.A. Multiply the difference (plus or minus) between expected and obtained number of words by multiplier corresponding to observed C.A.

2. For 4 vocabularies, combined scores are adjusted to a standard deviation of 15, which is that shown by boys of about 10 years 8 months and by girls of about 13 years 0 months. Mean of zero can be changed to 100 by adding 100 algebraically to the adjusted score.

3. For M.H.A. + M.H.B. scores are adjusted to a standard deviation of 5.7151, which is the root mean variance of girls in the experimental sample, and corresponds to a C.A. of about 9 years 0 months for boys and about 9 years 9 months for girls.

4. For T.M.L. the scores are adjusted to a standard deviation of 3.4357, which is the root mean variance of girls as before.

5. For W.I.S.C. the scores are adjusted to a standard deviation of 2.9851, which is the root mean variance of girls as before.

TABLE VIII. NORMS ; EXPECTED SCORES AND MULTIPLIERS

T.M.L.			W.I.S.C.			Exp'd No. of Words
Boys	C.A.	Girls	Boys	C.A.	Girls	
		5.0		5.0- 5.5	5.0- 5.5	8
					5.6- 5.11	9
5.0 - 5.4		5.1 - 5.8		5.6 - 6.0	6.0- 6.6	10
5.5 - 6.1		5.9 - 6.5		6.1 - 6.6	6.7- 7.1	11
6.2 - 6.9		6.6 - 7.2		6.7 - 7.1	7.2- 7.8	12
6.10- 7.5		7.3 - 7.10		7.2 - 7.9	7.9- 8.4	13
7.6 - 8.1		7.11- 8.6		7.10- 8.4	8.5- 9.0	14
8.2 - 8.9		8.7 - 9.2		8.5 - 9.0	9.1- 9.8	15
8.10- 9.5		9.3 - 9.10		9.1 - 9.8	9.9- 10.5	16
9.6 -10.0		9.11-10.6		9.9 -10.5	10.6-11.3	17
10.1 -10.7		10.7 -11.1		10.6 -11.1	11.4-12.2	18
10.8 -11.2		11.2 -11.9		11.2 -11.11	12.3-13.3	19
11.3 -11.9		11.10-12.5		12.0 -12.8	13.4-14.5	20
11.10-12.4		12.6 -13.0		12.9 -13.7	14.6-14.11	21
12.5 -12.10		13.1 -13.7		13.8 -14.6		22
12.11-13.5		13.8 -14.2		14.7 -14.11		23
13.6 -13.11		14.3 -14.9				
14.0 -14.6		14.10-14.11				
14.7 -14.11						

MULTIPLIERS

C.A.	T.M.L.		W.I.S.C.	
	Boys	Girls	Boys	Girls
5.0- 5.5	} 2.03	} 2.00	} 1.31	1.31
5.6- 5.11				1.31
6.0- 6.5				1.28
6.6- 6.11				1.25
7.0- 7.5	1.99		1.30	1.21
7.6- 7.11	1.65	2.00	1.21	1.17
8.0- 8.5	1.38	1.74	1.12	1.13
8.6- 8.11	1.18	1.46	1.03	1.08
9.0- 9.5	1.03	1.26	0.95	1.04
9.6- 9.11	0.92	1.10	0.89	1.00
10.0-10.5	0.84	0.99	0.83	0.97
10.6-10.11	0.78	0.90	0.79	0.93
11.0-11.5	0.73	0.83	0.75	0.91
11.6-11.11	0.69	0.78	0.72	0.88
12.0-12.5	0.66	0.73	0.70	0.86
12.6-12.11	0.64	0.70	0.68	0.84
13.0-13.5	0.62	0.68	0.67	0.82
13.6-13.11	0.62	0.66	0.67	0.81
14.0-14.5	0.62	0.65	0.68	0.80
14.6-14.11	0.62	0.64	0.69	0.80

M.H.A. + M.H.B.							
Boys	C.A.	Girls	Exp'd No. of Words	Boys	C.A.	Girls	Exp'd No. of Words
		5.0 - 5.1	4				
		5.2 - 5.4	5				
5.0 - 5.3		5.5 - 5.7	6	10.7 -10.10		*10.11-11.3	25
5.4 - 5.6		5.8 - 5.10	7	10.11-11.2		11.4 -11.7	26
5.7 - 5.9		5.11- 6.1	8	11.3 -11.6		11.8 -11.11	27
5.10- 6.1		6.2 - 6.4	9	11.7 -11.10		12.0 -12.3	28
				11.11-12.2		12.4 -12.8	29
6.2 - 6.4		6.5 - 6.8	10	12.3 -12.6		12.9 -13.1	30
6.5 - 6.7		6.9 - 6.11	11	12.7 -12.10		13.2 -13.5	31
6.8 - 6.11		7.0 - 7.2	12	12.11-13.2		13.6 -13.10	32
7.0 - 7.2		7.3 - 7.6	13	13.3 - 13.7		13.11-14.3	33
7.3 - 7.6		7.7 - 7.9	14	13.8 -13.11		14.4 -14.8	34
7.7 - 7.9		7.10- 8.1	15	14.0 -14.3		14.9 -14.11	35
7.10- 8.1		8.2 - 8.4	16	14.4 -14.8			36
8.2 - 8.4		8.5 - 8.8	17	14.9 -14.11			37
8.5 - 8.8		8.9 - 8.11	18				
8.9 - 8.11		9.0 - 9.3	19				
9.0 - 9.3		9.4 - 9.7	20				
9.4 - 9.7		9.8 - 9.11	21				
9.8 - 9.11		10.0 -10.2	22				
10.0 -10.2		10.3 -10.6	23				
10.3 -10.6		10.7 -10.10	24				

MULTIPLIERS

M.H.A. + M.H.B.					
C.A.	Boys	Girls	C.A.	Boys	Girls
5.0-5.5	} 1.63	1.90	10.0-10.5	0.84	0.97
5.6-5.11		1.73	10.6-10.11	0.79	0.92
6.0-6.5		1.59	11.0-11.5	0.75	0.88
6.6-6.11		1.47	11.6-11.11	0.72	0.85
7.0-7.5		1.36	12.0-12.5	0.69	0.82
7.6-7.11	1.24	1.27	12.6-12.11	0.67	0.79
8.0-8.5	1.13	1.20	13.0-13.5	0.65	0.76
8.6-8.11	1.04	1.13	13.6-13.11	0.64	0.74
9.0-9.5	0.96	1.07	14.0-14.5	0.63	0.72
9.6-9.11	0.90	1.01	14.6-14.11	0.62	0.69

FOUR VOCABULARIES COMBINED

Boys	C.A.	Girls	Exp'd No. of Words	Boys	C.A.	Girls	Exp'd No. of Words
				9.8		10.1 -10.2	50
		5.0	16	9.9 - 9.10		10.3 -10.4	51
		5.1 - 5.2	17	9.11-10.0		10.5 -10.6	52
		5.3	18	10.1 -10.2		10.7 -10.8	53
5.0		5.4 - 5.5	19	10.3 -10.4		10.9 -10.10	54
5.1 -5.2		5.6 - 5.7	20	10.5 -10.6		10.11-11.0	55
5.3 -5.4		5.8 -	21	10.7 -10.8		11.1 -11.2	56
5.5		5.9 - 5.10	22	10.9 -10.10		11.3 -11.5	57
5.6 -5.7		5.11- 6.0	23	10.11-11.0		11.6 -11.7	58
5.8 -5.9		6.1	24	11.1 -11.2		11.8 -11.9	59
5.10-5.11		6.2 - 6.3	25	11.3 -11.4		11.10-11.11	60
6.0		6.4 - 6.5	26	11.5 -11.6		12.0 -12.1	61
6.1 -6.2		6.6 - 6.7	27	11.7 -11.8		12.2 -12.4	62
6.3 -6.4		6.8	28	11.9 -11.10		12.5 -12.6	63
6.5 -6.6		6.9 - 6.10	29	11.11-12.0		12.7 -12.8	64
6.7		6.11- 7.0	30	12.1 -12.2		12.9 -12.10	65
6.8 -6.9		7.1 - 7.2	31	12.3 -12.4		12.11-13.1	66
6.10-6.11		7.3 - 7.4	32	12.5 -12.6		13.2 -13.3	67
7.0 -7.1		7.5	33	12.7 -12.8		13.4 -13.5	68
7.2 -7.3		7.6 - 7.7	34	12.9 -12.10		13.6 -13.8	69
7.4 -7.5		7.8 - 7.9	35	12.11-13.0		13.9 -13.10	70
7.6		7.10- 7.11	36	13.1 -13.2		13.11-14.1	71
7.7 -7.8		8.0 - 8.1	37	13.3 -13.4		14.2 -14.3	72
7.9 -7.10		8.2 - 8.3	38	13.5 -13.6		14.4 -14.6	73
7.11-8.0		8.4	39	13.7 -13.8		14.7 -14.8	74
8.1 -8.2		8.5 - 8.6	40	13.9 -13.10		14.9 -14.11	75
8.3 -8.4		8.7 - 8.8	41	13.11-14.0			76
8.5		8.9 - 8.10	42	14.1 -14.2			77
8.6 -8.7		8.11- 9.0	43	14.3 -14.4			78
8.8 -8.9		9.1 - 9.2	44	14.5 -14.7			79
8.10-8.11		9.3 - 9.4	45	14.8 -14.9			80
9.0 -9.1		9.5 - 9.6	46	14.10-14.11			81
9.2 -9.3		9.7 - 9.8	47				
9.4 -9.5		9.9 - 9.10	48				
9.6 -9.7		9.11-10.0	49				

MULTIPLIERS: FOUR VOCABULARIES COMBINED

C.A.	Boys	Girls
5.0- 5.5	2.19	2.28
5.6- 5.11		2.27
6.0- 6.5		2.19
6.6- 6.11		2.07
7.0- 7.5	1.94	1.93
7.6- 7.11	1.73	1.79
8.0- 8.5	1.54	1.66
8.6- 8.11	1.39	1.55
9.0- 9.5	1.26	1.45
9.6- 9.11	1.16	1.36
10.0-10.5	1.07	1.28
10.6-10.11	1.01	1.21
11.0-11.5	0.95	1.15
11.6-11.11	0.91	1.10
12.0-12.5	0.87	1.06
12.6-12.11	0.85	1.02
13.0-13.5	0.83	0.98
13.6-13.11	0.81	0.96
14.0-14.5	0.81	0.93
14.6-14.11	0.81	0.91

REFERENCES

- [1] DUNSDON, M. I., and ROBERTS, J. A. F. (1953). 'The relation of the Terman-Merrill vocabulary test to mental age in a sample of English children.' *Brit. J. Stat. Psychol.*, VI, 61-70.
- [2] DUNSDON, M. I., and ROBERTS, J. A. F. (1955). 'A study of the performance of 2,000 children on four vocabulary tests.' *Brit. Stat. J. Psychol.*, VIII, 3-15.
- [3] HAMILTON, M., PICKERING, G. W., ROBERTS, J. A. F., and SOWERBY, G. S. G. (1954). 'The aetiology of essential hypertension: (2) Scores for arterial blood pressures adjusted for differences in age and sex.' *Clin. Sci.*, XIII, 37-49.
- [4] RAVEN, J. C. (1948). *The Mill Hill Vocabulary Scale*. London: H. K. Lewis.
- [5] ROBERTS, J. A. F., NORMAN, R. M., and GRIFFITHS, R. (1935). 'Studies on a child Population: (i) definition of the sample method of ascertainment, and analysis of the results of a group intelligence test.' *Annals of Eugenics*, VI, 319-338.
- [6] ROBERTS, J. A. F., and MELLONE, M. A. (1952). 'On the adjustment of Terman-Merrill I.Q.s to secure comparability at different ages.' *Brit. Stat. J. Psychol.*, V, 65-79.
- [7] ROBERTS, J. A. F. (1953). 'The use of regressions involving variances of dependent variates for calculating age corrected scores.' *Biometrics*, IX, 267.
- [8] Terman, L. M., and MERRILL, M. A. (1937). *Measuring Intelligence*. London: Harrap.
- [9] WECHSLER, D. (1949). *Intelligence Scale for Children*. New York: The Psychological Corporation.
- [10] YATES, A. J. (1956). 'The use of vocabulary in the measurement of intellectual deterioration: A review.' *J. Ment. Sci.*, CII, 409-440.

PSYCHOLOGICAL FACTORS: THEIR NUMBER, NATURE, AND IDENTIFICATION

By G. BERNYER

Centre National de la Recherche Scientifique, Paris

Summary

The following paper presents a brief survey of current factorial procedures, with a discussion of their chief advantages and limitations. It is intended primarily for research workers in general psychology who desire to analyse the factors involved in their data without making a detailed study of all available procedures. An attempt is made to determine the commoner faults, either of theory or of practice, which have rendered so many recent factorial investigations contradictory or inconclusive; and a number of defects are noted in existing techniques which appear to call for further investigation by the mathematical theorist. The main conclusion is that, like other statistical procedures, factor analysis is merely an adjunct to psychological research, not a method sufficient in itself¹.

THE CHOICE OF PROCEDURES

During the past fifty years, factor analysis has undergone a rapid development and provoked a series of perplexing controversies. The literature, even if we include only publications dealing with factorial procedures and disregard those summarizing actual results, is now so vast that the investigator finds himself confronted by a formidable mass of contributions, which recommend widely different techniques and often lead to conclusions that are entirely contradictory. The psychologist who does not wish to embark on a comprehensive study of the whole subject, and wants merely to select the method most appropriate to his own immediate problem, is faced with a bewildering task if he is to appreciate the technical difficulties, the psychological implications, and the special limitations of the particular procedure that he has decided to apply. It may therefore be helpful to attempt a brief review of the practical questions encountered by the ordinary worker who desires to choose and apply some suitable technique. I shall confine my remarks chiefly to investigations involving tests of ability or aptitude, and shall deal chiefly with the two methods that are most frequently used by the psychologist, namely, the method of general and group factor analysis on the one hand, and the method of multifactorial analysis (including rotation to simple structure) on the other.

¹ *Editorial Note*—Both the author and the editor are indebted to the authorities of the *Centre National de la Recherche Scientifique* for permission to publish this paper here. The French version was originally presented at the International Colloquium on Factor Analysis, arranged by the *Centre* and published, with a full account of the discussion that followed, in the Report of their proceedings (Paris, 1955). Many requests have already been received for a copy to be available in English. We are grateful to Mlle. G. Régnier for undertaking the translation, and to Dr. Charlotte Banks for revising the manuscript. Students and others can obtain reprints at cost price by writing to the Assistant Editor, Dr. Charlotte Banks, University College, London.

FUNDAMENTAL ASSUMPTIONS

Both procedures start with the broad hypothesis that the marks obtained by a particular examinee in a particular test may, to a first approximation, be treated as a linear function of certain hypothetical variables or 'factors'; and both recognize two types of factor, namely, (a) 'common factors', i.e. those that appear either in all the tests (the general factor) or in some of the tests (group factors), and (b) 'specific factors', i.e. those peculiar to some single test alone. All such factors, except perhaps those that are invoked in the later stages of multifactorial analysis, are assumed to be statistically independent.

Here it is too often forgotten that, even when we restrict ourselves to this apparently simple hypothesis, the problem of determining the factors still possesses an infinity of solutions. To meet this difficulty and limit the number of solutions available, most writers commonly invoke a 'principle of economy', very much the same in both procedures. This principle requires us to extract a minimum number of common factors. By thus minimizing the common factor variance, however, we maximize the specific factor variance. Moreover, if each of the two alternative procedures take as their point of departure one and the same bipolar analysis, the results turn out to be statistically equivalent; we can pass from one set to the other by an appropriate rotation. On the other hand, the factors actually furnished by the two methods of analysis differ widely from each other in consequence of the further postulates which each imposes in order to reach the final solution.

ANALYSIS BY GENERAL AND GROUP FACTORS

The 'two-factor theory' of Spearman [1] sought to interpret the marks obtained in a test by each examinee by means of a single general factor only, together with a factor specific to the test. This seemed to provide a solution of the factorial problem which was attractive at once by its neatness and by its simplicity. But, as soon as it was fully realized that this simple assumption fails to fit the psychological data [2], and that a single general factor, g , is insufficient to explain all the correlations between the tests, it became necessary to reintroduce group factors as well [3]. Spearman himself for long refused to admit the need for such supplementary factors; and, when at length he agreed to adopt them, he again applied a strict principle of economy, and insisted that the group factors should be as few as possible and that their existence and statistical significance should first be fully established.

With a view to determining their number and extent, he himself relied on the examination of tetrad differences [4]. In its practical application, however, this procedure presents a number of grave difficulties. It is extremely laborious; it still does not always lead to a unique factorial solution; and the sampling errors for the tetrad differences can only be calculated by the aid of formulae which seem in point of fact to be somewhat crude approximations. In practice, therefore, the employment of this procedure has remained confined to small tables containing only a few tests and including only a single group factor—a limitation that has not always been noticed.

An alternative method, more recent, which has been largely overlooked by psychologists, could be substituted with great advantage for that employed by Spearman. This is the method proposed by P. Delaporte [5]. It permits much larger batteries to be analysed, and allows us to use tables giving the confidence intervals for the ratio between the correlation coefficients observed. It thus permits the investigator to establish the factorial pattern quite quickly and without detailed calculation. However, in its practical application a few minor questions still remain to be solved: in certain cases it is by no means easy to decide whether the principle of economy should be applied to the number or to the extent of the factors; and consequently a subjective choice, which takes into account the actual content of the test, will then be necessary.

Very few psychologists have in point of fact continued to follow the line proposed by Spearman. Burt has reviewed the reasons which led him to criticize this simplified method of analysis [6]. With the somewhat drastic principle of economy advocated by Spearman the chief difficulty is that the general factor which is actually extracted can vary widely as a result of the modifications that may be introduced in the composition of the test-battery. In the majority of cases its effect is to suppress the broadest group factor, which gets incorporated into the general factor, and as a result imparts to the modified general factor an additional content which should properly be assigned to the group factor, for example, a verbal content or a spatial content according to the nature of the tests employed. As a result the general factors so obtained cease to be comparable from one battery to another.

To overcome this difficulty, several investigators have adopted the bi-factorial method of Holzinger [7] which interprets the intercorrelations of the tests in terms of a general factor, g , and a number of group factors which are mutually exclusive. However, the procedures proposed for deciding how the tests are to be grouped, for calculating the factor saturations, and for testing the statistical significance of the residual correlations left after the common factors have been extracted, appear for the most part to be rather crude.

In a recent paper on group factor analysis [8] Burt has furnished a method of great value for those psychologists who wish to carry out a group factor analysis on a more precise basis. Readers will there find, clearly summarized, a critical study of the various procedures most commonly used for establishing the factor pattern (subjective classification, the B coefficient of Holzinger, Tryon's method of profiles, and the device of intercolumnar correlations), together with an account of what Burt calls the 'method of summation' for calculating the factor saturations illustrated by worked examples. Moreover, the procedure described indicates how a preliminary bipolar analysis may be used with advantage to determine the best grouping of the tests, and also allows the investigator to calculate factor saturations even when there is a partial overlap between certain of the group factors.

SUBDIVIDED FACTORS

It is likewise to Burt that we owe an important extension of the group factor procedure, namely, the method of 'subdivided factors' [9]. After describing

the hypothesis put forward by McDougall and other writers according to which the structure of the mind is believed to be essentially 'hierarchical', Burt shows that a divergent scheme of classification often proves to be applicable to psychological data, and that, when a table of correlations fulfils certain conditions (in particular when the residual cross-correlations remaining after a bipolar factor has been extracted are non-significant), it is possible to demonstrate over and above the general factor certain broad group factors (e.g. an intellectual factor and a practical factor) which in turn subdivide themselves into narrower group factors (e.g. a verbal factor, a numerical factor, and the like, subsumed under the heading of the intellectual group factor) and such that these in turn can be further subdivided into factors which are narrower still.

Many British authors (e.g. Vernon) have successfully used factor patterns of this type. The preliminary bipolar analysis serves as a point of departure for the calculation of the subdivided group factors. And with this mode of analysis not only the arrangement, but also the number of factors extracted (which is somewhat greater than that of the corresponding bipolar factors obtained with the classical procedures) is automatically determined by the number of factors extracted in the preliminary bipolar analysis. Thus, if we take a common instance, namely, a bipolar analysis involving four factors—one general and positive, one bipolar dividing this into two, and two more subdividing its positive and negative sections into two subsections each, with cross-correlations of approximately zero, we find that, when expressed in terms of a group-factor classification with positive saturations only, this yields as many as seven subdivided factors—one general and a pair of broad group factors, each subdividing into a different pair of narrower factors.

Although group factor analysis has been used fairly widely in psychology, the procedure has its limitations. In the first place, if the preliminary method of calculation is based on the method of simple summation, which is ordinarily substituted for the more exact technique of weighted summation, then the more rigorous tests for statistical significance, which are appropriate to weighted summation, are no longer strictly valid, and must be regarded as approximate only. In the second place, group factor analysis, like every other factorial procedure, can only be applied to batteries of tests which have been carefully planned in view of the problem to be investigated. "The common idea", says Burt, "that factor analysis can be mechanically applied to any set of data that happens to be handy, and that factors can thus be 'discovered', is entirely misleading."

MULTIFACTORIAL ANALYSIS

Thurstone, in opposition to the theory put forward by Spearman, starts from the hypothesis that the inter-correlations of a battery of tests are to be explained, not in terms of a single general factor, but by a small number of common factors [10, 11]. As is well known, the proof that all the tetrad differences (i.e. minor determinants of order two) are equal to zero is the mathematician's

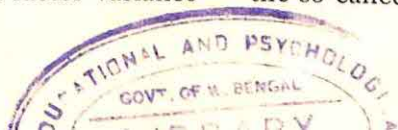
regular way of demonstrating that the rank of the matrix from which they are derived is one. Thurstone invokes the general form of this theorem. "In the multifactorial problem", he says, "we inquire at the outset what is the rank of the given table of intercorrelations; and this rank indicates the number of factors which have to be postulated in order to interpret the correlations experimentally obtained." However, in practice "the rank of a correlation matrix is always equal to its order, because of the accidental errors of sampling entering into each of its elements. Consequently, we are compelled to make use of an approximation to the rank of the matrix. We ask, in point of fact, whether it is possible to construct a theoretical correlation matrix of lower rank, which will only differ to a slight extent from the correlations experimentally observed. When such a correlation matrix can be found, then its rank gives the number of factors which have to be postulated to interpret the observed correlations" [12].

The method thus proposed by Thurstone for solving this problem appears to involve several difficulties, which I shall now briefly attempt to indicate.

COMMUNALITIES

The first question to be settled is the size of what are called the 'communalities', that is to say, the quantities which are to be entered in the principal diagonal of the correlation matrix in order to make its rank a minimum. Now, at the beginning of the analysis, the rank of a correlation matrix is unknown; and strictly speaking we can only secure estimates for the communalities by determining (as in the case of a correlation matrix of rank one) the lowest order of minor determinants which are equal to zero. Several working procedures, all of them more or less crude, have been suggested and discussed. But the majority of analysts appear content to adopt Thurstone's short cut, and, before extracting each factor, insert in each diagonal cell the largest correlation in the column. When the number of tests is large, the error thus introduced is comparatively unimportant. However, Burt and other English investigators prefer to carry out the estimation of the communalities or self-correlations by successive approximation. With smaller tables of correlations some investigators may be sufficiently interested to try the method of maximum likelihood described by Lawley, provided they are ready to face the rather laborious calculations involved. This yields efficient estimates, and incidentally provides a criterion for deciding the number of factors to be extracted [13 and refs.].

The next problem consists in computing from the table of correlations as thus completed the various common factors which will account for the observed coefficients within the margin allowed by the fluctuations of sampling. Thurstone seems now [14] to view with more favour than before [11] the method of 'principal components' as described by Hotelling: this is the procedure with which British writers are more familiar as the 'method of principal axes', originally described by Karl Pearson [15]. With this procedure we obtain a series of orthogonal factors by "extracting at each stage that particular factor which will account for a maximum amount of the common factor variance"—the so-called method of



'weighted summation' [16]. The method of 'simple summation' described by Burt and the equivalent procedure called the 'centroid method' by Thurstone are quicker to apply, and furnish a reasonably good approximation to the principal components; the centroid method is that most commonly employed by American writers. All these procedures can readily be applied to practically any table of correlations. The chief difficulty that remains is to determine at what point the analysis must be stopped.

NUMBER OF FACTORS

How many factors should be extracted has been a question on which numerous inquiries have been published, and is still far from being completely solved. Many criteria have been put forward for testing the statistical significance of the residual correlations which remain after a given number of common factors have been extracted [17]. Of these the more exact demand lengthy calculations and moreover are not valid when the analysis has been carried out by the centroid procedure. The majority are based on more or less empirical rules, which in point of fact scarcely yield any more information than can be obtained by simple inspection without any calculations derived from the residuals. Indeed, in practice, many factorists employ none at all. Thurstone and his pupils apparently prefer to extract more factors than is strictly necessary, and then at the end of the analysis to discard those which are not susceptible of psychological interpretation. This lack of rigour is perhaps partly justified, since, as Thomson has indicated, the sampling errors can themselves produce common factors which will not be concentrated in the last centroid factor, but will be mingled with all such factors; and Burt has further insisted that "the discovery that, with the sample used, a given factor turns out to be non-significant cannot be taken to prove that it is non-existent".

Actually the aim of the psychologist is not so much to extract from a correlation matrix the minimum number of factors as to verify the existence of the factors that he has postulated, and if necessary to discover whether there may not be further groupings of the tests employed which have hitherto been overlooked. In order to judge whether the number of factors extracted is sufficient, he will in the end nearly always need to examine the particular results obtained. For example, in an analysis where three factors furnished non-significant residuals as judged by McNemar's criterion, I myself have been able to show, simply by examining the diagram obtained on plotting the test-vectors, that four factors at least would be required to group the tests correctly and achieve a simple structure [18].

The greatest difficulty that besets the use of factor analysis springs from the fact that the 'components' or 'mental factors' that such investigations continue to reveal do not form a small or sharply limited group. On the contrary, as the facts seem clearly to prove, the progress of factorial work leads to an ever increasing number. In a recent monograph dealing with the description of aptitude and achievement tests in terms of rotated factors [19], J. W. French has enumerated

well over sixty. Now in any given factorial investigation it is possible to extract only a very few fully significant factors. The actual number is unavoidably limited by the number of the subjects who have been tested. After a study of a series of different researches Burt considers that, as a rule, with a battery of about ten to twenty tests, at least twenty subjects are needed to establish one factor, at least fifty subjects to establish two, over a hundred to establish three, and as many as two or three hundred to establish four. Thus a factorial analysis can achieve a result which is at once statistically accurate and capable of a clear interpretation only when the number of factors aimed at has been scrupulously limited from the very outset.

SIMPLE STRUCTURE

The different methods of analysis which we have briefly enumerated—the method of simple summation or centroid analysis, the method of principal components or weighted summation, the method of maximum likelihood—yield factors which have negative loadings in certain tests and which often vary widely if the composition of the test-battery is changed. Such factors, so Thurstone contends, must be devoid of ‘scientific meaning’. They do not permit us to “interpret the various tests as functions of the mental aptitudes which those tests elicit”. In order to acquire a meaning, therefore, the factors should be subjected to rotations, which are continued until the factorial matrix has the largest possible number of values which are either zero or approximately zero. If the investigator succeeds by this device in obtaining at least one zero saturation in each row, and at least as many zeros in each column as there are factors, he is said to have achieved a ‘simple structure’. When such a structure has been secured, however, the group factors that emerge are no longer of necessity orthogonal to each other. Thurstone regards them as ‘primary factors’, and believes them to be invariant and susceptible of psychological interpretation [14].

In practice, the search for such a solution is not without difficulties of its own. No method that is absolutely automatic and objective has been devised for securing it. And the successive rotations which are needed require long and laborious trials. In practice, since there are no formulae for determining the errors in the factor saturations thus obtained, many analysts have arbitrarily adopted their own limit of significance; and indeed often seem a little too prone to consider that they have satisfied the criteria for a simple structure. By such procedures they manage to start from almost any kind of correlation table and reach a set of ‘primary factors’ whose stability and whose psychological interest are in point of fact far from being adequately established. Here assuredly there lies a great danger—one in fact on which Thurstone himself has insisted. “Factor analysis”, he says, “is a scientific method which has to be adapted to each problem: it is no mere statistical procedure; and it is not a routine device which one can fruitfully apply to any table whatsoever.” It would seem, then, that every factorial investigation ought to start with a plan of inquiry, and that a ‘simple structure’ can have no definite meaning unless it is reached as a confirmation of factors postulated at the outset.

THE IDENTIFICATION AND INVARIANCE OF FACTORS

By way of defending the superiority of their own particular method, the champions of different factorial procedures commonly assert that theirs will yield factors which can be most readily identified and which will remain the most stable when one changes either the battery of tests or the population tested. And in order to secure an objective basis for determining the identity of the factors thus extracted, a number of widely different procedures have been devised. The following seem to be the most satisfactory suggestions [20]: (i) when several wholly independent analyses have been carried out, to compare the amount contributed to the variance by corresponding factors in different analyses; (ii) when the experiments have been carried out on the same group of subjects, to calculate the correlations between the estimates for corresponding factors; (iii) when the experiments have been carried out with the same battery of tests, to compare the factor saturations for the several tests by means of an appropriate coefficient: for this purpose Burt has devised a 'coefficient of similarity' between the two corresponding sets of saturations, while Cattell prefers to use a 'marker criterion' or a 'coefficient for similarity of patterns'.

Hitherto these procedures have been but rarely used, and that for several reasons. Until quite recently there have been comparatively few factorial investigations whose results would permit their application: it is in fact rather exceptional for the factor measurements to be estimated in detail for each of the persons tested; and, even so, a knowledge of the reliability of the measurements would seem to be necessary before any trustworthy information can be deduced from their correlations. Moreover, it is seldom possible to compute even the simpler type of coefficient based on a comparison of the factor saturations, since, in their desire to carry out original work, the majority of authors appear reluctant to adopt batteries that have already been used in previous researches, and nearly always prefer to substitute new tests of their own. French observes that, in the comparative study that he has made of some 69 analyses, he has been obliged to limit himself to a subjective comparison of the tests because there were only four instances in which precisely the same batteries had been employed [19]. Moreover, the practical value of the different procedures still needs to be more fully demonstrated. Leyden [21], using data from a battery of tests administered to the same group of pupils at intervals of 1 to 3 years, has applied most of the foregoing criteria to test the stability of the general factor, and finds that they give very divergent results. On the whole, the so-called coefficient of similarity devised by Burt would seem to be the best for general use. However, such methods merely supply a general indication as to whether or not the factorial pattern exhibits any significant change, and should always be supplemented by a more detailed study of the material in order to take full account of all the complex issues involved. On applying one and the same battery to two populations at different levels, I myself found that, with Thurstone's procedure, the factorial composition of some of the tests appeared to change appreciably, and so modified the theoretical pattern, although for the majority of tests the saturations remained

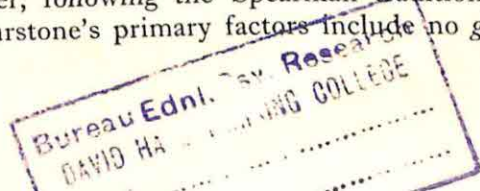
fairly stable, and thus at first sight suggested that the factors extracted in the two investigations were in fact identical.

Nevertheless, even when no rigorous methods are applicable for identifying and comparing factors, the factorist should not confine himself merely to a superficial examination of the factor saturations and of the psychological content of the tests. He should also take into consideration the general conditions of the experiment as a whole. The question how far the influence of the general factor, g , diminishes with age, while at the same time special aptitudes become more and more plainly differentiated, provides one of the most typical examples of the difficulties arising in interpreting factorial results. Numerous researches have been carried out on this problem, and have led to very discrepant and even contradictory conclusions. In a brief analysis of these various inquiries, Vernon [22] has shown clearly that alterations in the factorial schemes depend not only on the composition of the test batteries, but also on the degree of heterogeneity or selection in the populations tested and on their educational or vocational training: he emphasizes the fact that, even if random samples of the population could be tested at different age levels, the content of the tests would also have to be adapted to the different ages, in order not to introduce modifications into the variance of the first factor. It would seem, therefore, that the influence of age presents a problem which is peculiarly difficult to solve.

THE NATURE OF THE FACTORS

When considering the problems raised when we attempt to determine the nature of the factors, it is well to bear in mind the view expressed by Thomson. "I do not believe", he writes, "that the factors arrived at by minimizing the number of common factors and maximizing the part played by specifics can be in any sense realities, either in a physiological or a psychological sense. I think that the causes of ability to do well in this, that, or the other test are innumerable and small, some of them genetic, some of them acquired, and that the correlations or resemblances between test-lists are due to some of these causes being helpful in both tests, which overlap in the sample they draw on from these innumerable little causes, which I would not dream of calling 'factors', and which are different in different persons. It remains true, however, that factors form a vivid and convenient description of any battery of tests, a description we would be silly to forgo, but also a description we must keep in its place and not raise to the status of a causal explanation" [23]. All this appears to me decidedly helpful, because it is likely to induce the psychologist to recognize the difficulties and the limitations, as well as the possibilities, of factorial methods.

Moreover, if factors are thought of as possessing essentially a descriptive role, then the divergences between the different factorial techniques will no longer seem irreconcilable. "The big difference", as Thomson himself points out, "between Thurstone's American school of thought and the view of a majority of British psychologists is that the latter, following the Spearman tradition, cling to a general factor g , whereas Thurstone's primary factors include no g .



The difference is not fundamental except to those who reify their factors and look upon them as actual entities, possibly with physiological bases such as a certain brain centre, a certain endocrine gland, or the like. Others, who look upon factors only as convenient descriptive parameters, realize that the two descriptions, with or without a *g*, are equally capable of accurately specifying a test or a man."

From a theoretical standpoint, the antagonism between the two schools is now tending to disappear. In adopting oblique factors, Thurstone has indirectly recognized the existence of a general factor; and he has since explicitly reintroduced such a factor in his recent attempt to re-analyse his centroid factors in terms of what he calls 'second order factors' [14]. Possibly the differences between these alternative types of analysis might have greater practical significance, were the factors themselves estimated in detail for each of the subjects tested. In practice, however, psychologists rarely make such estimates, and appear content to employ batteries containing groups of tests that are presumed to correspond with certain important factors, e.g. the verbal factor or the spatial factor; as a result, no matter what method of analysis may have been relied on at the outset when constructing such batteries, the groups of tests which they contain seem to all intents and purposes to be very much the same.

The view that the relations between tests are to be interpreted in terms of a large number of small causes suggests a reason why factor analysis does not help us to 'discover factors' and why it has proved so difficult in the course of such investigations to secure a small number of factors that are really stable. Beyond a doubt, no psychologist at the present day would pretend that it is sufficient to apply a large variety of tests to a large number of persons in the hope of determining, in a single research, all the 'factors of the mind'. Nevertheless, many investigators still embark on factorial inquiries without clearly realizing that the factors extracted will depend largely on the choice of the tests which they put into their batteries, and that factorial analysis cannot be safely employed except in strictly limited types of experiment, namely, those that have been deliberately designed with the needs of a factorial study in view. It is therefore only valid when, on the basis of some initial psychological hypothesis, a provisional classification of the data has been explicitly laid down at the outset. It is this constant neglect of the intrinsic limitations of factorial procedures that appears to be the main reason why so many arduous investigations now seem to have been entirely wasted.

CONCLUSIONS

The main conclusions that emerge from this broad review may be briefly summed up as follows. Factorial analysis is not a mode of psychological research which is sufficient in itself. Unaided, such a method, as Burt has said [24], does not permit a determination of causal influences: these can only be discovered "by means of experimentation, of controlled observation, and of introspection". Factor analysis is, in his view, essentially a statistical technique which allows us to verify and specify with greater precision a hypothetical classification, and which may, in certain cases, tend to modify or even to suggest such a classification.

It is in short a mathematical test like calculating Student's t , or chi-squared, or an analysis of variance. Confined to this role it provides the psychologist with a supplementary aid of great value which can be used *in the course of*, but not as the sole basis of, a psychological research. It is to be hoped that a further development of factorial techniques will render such verificatory tests more exact and more effective.

REFERENCES

- [1] SPEARMAN, C. (1914). The theory of two factors. *Psychol. Rev.*, XXI, 101-115.
- [2] BURT, C. (1909). Experimental tests of general intelligence. *Brit. J. Psychol.*, III, 94-177.
- [3] BOARD OF EDUCATION (1924). *Report on Psychological Tests of Educable Capacity*. London: H.M. Stationery Office.
- [4] SPEARMAN, C. (1927). *The Abilities of Man*. London: Macmillan.
- [5] DELAPORTE, P. (1947). Prolongement de la méthode d'analyse factorielle de Spearman en utilisant la statistique mathématique. *Biotypologie*, IX, 45-59.
- [6] BURT, C. (1949). The two-factor theory. *Brit. J. Psychol. (Stat. Sect.)*, II, 151-179.
- [7] HOLZINGER, K. J., and HARMAN, H. H. (1941). *Factor Analysis: a Synthesis of Factorial Methods*. Chicago: University of Chicago Press.
- [8] BURT, C. (1950). Group factor analysis. *Brit. J. Psychol. (Stat. Sect.)*, III, 40-75.
- [9] BURT, C. (1949). Subdivided factors. *Brit. J. Psychol. (Stat. Sect.)*, II, 41-63.
- [10] THURSTONE, L. L. (1931). Multiple factor analysis. *Psychol. Rev.*, XXXVIII, 406-427.
- [11] THURSTONE, L. L. (1935). *The Vectors of Mind*. Chicago: University of Chicago Press.
- [12] THURSTONE, L. L. (1951). L'analyse factorielle: méthode scientifique. *Année psychologique*, L, 61-75.
- [13] EMMETT, W. G. (1949). Factor analysis by Lawley's method of maximum likelihood. *Brit. J. Psychol. (Stat. Sect.)*, II, 90-97.
- [14] THURSTONE, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.
- [15] PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, II, 6 Ser., 559-572.
- [16] BURT, C. (1949). Alternative methods of factor analysis and their relations to Pearson's method of 'principal axes'. *Brit. J. Psychol. (Stat. Sect.)*, II, 98-121.
- [17] BURT, C. (1952). Tests of significance in factor analysis. *Brit. J. Psychol. (Stat. Sect.)*, V, 109-133.
- [18] BERNYER, G. (1956). Analyses factorielles de deux batteries de tests appliquées aux candidats d'une école de mécanique. *Biotypologie*, XVII (in the press).
- [19] FRENCH, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs*, No. 5. Chicago: University of Chicago Press.
- [20] BARLOW, J. A., and BURT, C. (1954). The identification of factors from different experiments. *Brit. J. Stat. Psychol.*, VII, 52-56.
- [21] LEYDEN, T. (1953). The identification and invariance of factors. *Brit. J. Stat. Psychol.*, VI, 119.
- [22] VERNON, P. E. (1950). *The Structure of Human Abilities*. London: Methuen.
- [23] THOMSON, G. H. (1951). Factor analysis, its hopes and dangers. *Proc. XIIIth Internat. Cong. Psychol.* Stockholm.
- [24] BURT, C. (1947). L'analyse factorielle dans la psychologie anglaise. *Biotypologie*, IX, 7-44.

THE COMPARISON OF FREQUENCIES IN MATCHED SAMPLES

By ALAN STUART

Research Techniques Unit, London School of Economics

Abstract

For purposes of statistical comparison, matched samples are in general more accurate than unmatched. In testing the statistical significance of the results, it is essential to take into account the gain in accuracy, due to the matching. The following paper indicates how this can be done in the case of attributes.

1. INTRODUCTION

Psychologists and others often seek to eliminate the effects of unwanted sources of variation on comparisons between two groups by 'matching' individuals, one from each group, for values of the variables to be eliminated. If a quantitative variable is being investigated, differences between matched individuals may be tested by a Student's t test or some analogous distribution-free procedure. A similar test is required when an attribute is being investigated, but does not seem to have been made available. It turns out that such a test can be very simply derived (almost entirely by reference to the ordinary 2×2 table), and that it yields a test statistic whose distribution is approximately normal. A numerical example will be given in the last section.

2. THE STANDARD TEST

Let us write the 2×2 table in the usual form as follows:

	Number possessing attribute	Number not possessing attribute	Sample size
First sample	a	b	$a+b$
Second sample	c	d	$c+d$
Combined samples	$a+c$	$b+d$	$a+b+c+d=n$

The standard test for the equality of the proportions in the two groups from which random samples were drawn is made by computing

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad (1)$$

If all the cell frequencies are of reasonable size (say 5 or more), this will be distributed with a close approximation to the chi-squared distribution with 1 degree of freedom.

Another way of looking at the test is to remember that a chi-squared variate with 1 degree of freedom is the square of a standardized normal variate. Thus

$$\chi = (ad - bc) \sqrt{\left\{ \frac{n}{(a+b)(a+c)(b+d)(c+d)} \right\}}$$

is approximately such a variate. Let us re-arrange (2) to give the algebraically identical equation:

$$\chi = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\sqrt{\left\{ \frac{a+c}{n} \cdot \frac{b+d}{n} \cdot \left(\frac{1}{a+b} + \frac{1}{c+d} \right) \right\}}}. \quad (3)$$

We then see that χ is simply the ordinary test for the difference between the observed proportions $a/(a+b)$ and $c/(c+d)$, the difference being tested against its standard error estimated from the combined samples.

If the samples are of equal size, namely, $a+b = c+d = \frac{1}{2}n$, then (3) becomes

$$\chi = (a - c) \sqrt{\left\{ \frac{n}{(a+c)(b+d)} \right\}}. \quad (4)$$

3. THE EFFECT OF MATCHING

We now examine the effect of a matching procedure upon our test statistic. As before, we wish to test the difference in the numerator of χ in (3). On the hypothesis that there is no true difference between the matched groups, that numerator will have a mean value of zero, as in the unmatched case; but its standard error will be reduced by the matching process, which has the effect of a stratification upon a random sample. In the denominator of (3) the factor

$$\frac{a+c}{n} \cdot \frac{b+d}{n} = \frac{a+c}{n} \left(1 - \frac{a+c}{n} \right), \quad (5)$$

which is an estimate of the ordinary binomial (Bernoullian) variance, should therefore be replaced by the unbiased within-strata (Poissonian) estimate,

$$2n^{-1} \sum_{i=1}^{n/2} 2\left\{ \frac{1}{2}(x_i + y_i) \right\} \left\{ 1 - \frac{1}{2}(x_i + y_i) \right\}, \quad (6)$$

where x_i, y_i refer to the first and second samples respectively, and take the value 1 when the attribute is present in an individual and 0 when it is not. Equation (6) simplifies to

$$\frac{2}{n} \sum \{ (x_i + y_i) - \frac{1}{2}(x_i - y_i)^2 \}; \quad (7)$$

and this becomes

$$\frac{2}{n} \{ (a+c) - \frac{1}{2} \sum (x_i^2 + y_i^2 + 2x_i y_i) \}, \quad (8)$$

since $\sum x_i = a$ and $\sum y_i = c$. Moreover, since x_i and $y_i = 0$ or 1 for all i , we may further simplify (8) to

$$\frac{2}{n} \{ (a+c) - \frac{1}{2} \sum (x_i + y_i + 2x_i y_i) \} = \frac{1}{n} \{ (a+c) - 2S \}, \quad (9)$$

where we have written $S = \sum x_i y_i$. Here S is the number of matched pairs in which *both* individuals possess the attribute. Substituting (9) for (5) in (3), and putting $(a+b) = (c+d) = \frac{1}{2}n$, we obtain

$$Y = \frac{a-c}{\sqrt{(a+c-2S)}} \quad (10)$$

as our standardized normal variate for the matched case. In testing for significance Y is referred to the usual critical values of the normal distribution.

Unlike (4), its analogue in the unmatched case, (10) does not explicitly contain n , the effect of total sample size being expressed by the other terms. Moreover—though this is not at once evident from inspection—(10) is symmetric as between a and c on the one hand and b and d on the other. If we transpose the columns of the 2×2 table, we get exactly the same result, as we should for any reasonable test.

4. SENSITIVITY OF THE TEST

Whether (10) provides a sensitive test in any particular situation will depend upon whether the matching procedure has been efficacious. Suppose, for instance, that the matching makes no difference to the joint occurrence of the attribute, i.e. suppose that the probability that an individual whose matched 'twin' possesses the attribute will himself also possess it, is exactly the same as the overall probability of possessing the attribute, namely, $(a+c)/n$: then the expected number of matched pairs, each possessing the attribute, will be

$$\mathcal{E}(S) = \frac{1}{2}n\{(a+c)/n\}^2 = (a+c)^2/2n. \quad (11)$$

If in (10) we substitute (11) for S , we obtain for Y precisely the value of χ in (4). That is just what we should expect.

Clearly, if S tends to be larger than the value given by (11), then (10) will be increased and the test made more sensitive. Conversely, if it tends to be smaller, we may actually lose efficiency by matching, since the individuals of a matched pair may behave more differently than if unmatched. This is not, of course, a contingency likely to arise very often in practice; but it is worth mentioning to underline the fact that the benefits of matching are not automatic. In normal circumstances, where it is reasonably certain that matching has done its work properly, it would consequently be reasonable to use (4) as an effective lower bound for (10) in cases where for some reason S cannot easily be computed. This will be less wasteful of information than simply ignoring the term $2S$ in (10).

The extreme example of the gain in sensitivity which may accrue as the result of matching will occur when S takes its maximum value, which will be the lesser of a or c . In this case, (10) gives

$$|Y| = \sqrt{|a-c|},$$

which is significant at the 5 per cent level (± 1.96) whenever $|a-c| \geq 4$.

5. CORRECTION FOR CONTINUITY

In using the standard test statistic (1), we can improve the approximation of its distribution to the chi-squared distribution by applying Yates' correction for continuity. The effect of this is to replace the term $(ad - bc)^2$ in the numerator of (1) by $(|ad - bc| - \frac{1}{2}n)^2$. The same principle applies here; and in (10) the correction will have the effect of replacing $|a - c|$ by $|a - c| - \frac{1}{2}$.

6. EXAMPLE

The following illustration is based on fictitious data. 50 boys and 50 girls, matched in pairs by age, size of family, father's educational level, and number of rooms per person in the home, were studied in order to determine whether there is any sex-difference in performance at the 11 plus examination. The results were as follows:

Pupils	Obtained Grammar School place	Did not obtain Grammar School place
Boys	5	45
Girls	13	37

The value of S , the number of matched pairs in which both boy and girl obtained a grammar school place, was 4. We have, from (10),

$$Y = \frac{5 - 13}{\sqrt{\{5 + 13 - (2 \times 4)\}}} = -2.53.$$

Although the difference is small, it is fully significant at the 5 per cent level, and indeed almost at the 1 per cent level. The effect of the continuity correction would be to reduce Y to about -2.3 .

I am grateful to Mr. J. Durbin for suggesting several improvements in the presentation of the paper.

HEREDITY AND INTELLIGENCE: A REPLY TO CRITICISMS

By CYRIL BURT and MARGARET HOWARD
University College, London

Summary

An attempt is made to remove current misconceptions about the way in which individual differences in general intelligence appear to be inherited. Further support for a multifactorial form of the Mendelian theory is found in an examination (a) of the commoner objections, both methodological and statistical, and (b) of the chief alternative hypotheses, viz., (i) the theory of blending inheritance and (ii) the theory of exclusively or predominantly environmental determination.

I. THE COMMONER TYPES OF CRITICISM

Nature and Importance of Problem

In several earlier publications (Burt [2], [5] and refs.) one of us endeavoured to show how effectively the Mendelian principles of inheritance seemed to explain the observable distribution of mental differences, especially differences in what is commonly termed 'general intelligence'. During the past few years the problem of innate differences has come once again to the fore, particularly in connection with the allocation of pupils under the Education Act of 1944 and the Report of the Royal Commission on Population in 1949. In the meantime the psychological surveys carried out in London, Bath, Scotland, and elsewhere have furnished new data; while the recent advances in genetics, the fuller understanding of the mechanisms of heredity, and above all the improvements in biometric techniques, now make it possible to re-examine the basic issues with greater hope of success.

In the preceding issue of this Journal [9] we endeavoured to defend, on the basis of a more precise analysis, a theory previously put forward in general outline—namely, that innate differences in intelligence are due to Mendelian factors of two kinds: (i) major factors responsible for comparatively *large* deviations, usually of an abnormal kind, and (ii) multiple factors whose effects are *small, similar* and *cumulative*. However, the hypothesis that intelligence is inherited, both in the form that we had suggested and in various other versions, has of late been subjected to vigorous attack. In the main, the criticisms appear to fall into three more or less overlapping groups: first, those directed against the psychological and biological concepts in terms of which the hypothesis is expressed; secondly, those arising out of the abstract methodology of such inquiries; and thirdly, those concerned with the statistical techniques. As a rule, each critic singles out merely those particular aspects that happen to interest him, and ignores the other lines of argument. Nothing like a systematic survey of the various difficulties has so far been attempted. Some such review

seems urgently needed, since a study of the commoner objections and misunderstandings may help to set both the theory and the reasons for it in a sharper light, and may at the same time obviate a number of practical proposals which appear doomed to disappointment.

II. GENERAL MISCONCEPTIONS

The Meaning of Heredity

The most frequent type of criticism springs from misunderstandings about the exact nature of the issues involved and of the hypothesis put forward. Two distinct questions are constantly confused: (i) in what way, if at all, are genetic differences in ability transmitted? (ii) how accurately do existing tests or other methods of estimation enable us to assess such differences?

In the first place, both critics and supporters of the doctrine of mental heredity seem as a rule to hold quite obsolete and misleading notions about the way biological characteristics are transmitted. Tacitly if not explicitly, 'heredity' is assumed to mean 'the tendency of like to beget like' (the definition quoted by one critic from the *Oxford English Dictionary*). Consequently, the arguments for inheritance are supposed to consist in demonstrating resemblances between the parent and his children. When the two parents differ, the child is supposed to consist in a blend of both—rather on the lines of what the Arabian sage declared would happen if you mated a white elephant with a black giraffe, or the views that led older biologists to predict such hybrids as the centaur, the mermaid, and the minotaur, and induced even Aristotle to describe the giraffe as the progeny of a camel and a leopard.

The approach of the modern geneticist is the reverse of all this. As he sees it, the real problem is rather to explain why in so many instances 'like begets unlike'. And the only satisfactory answer is that supplied by the 'mosaic theory of inheritance' which we owe to Mendel:

An atomic mechanism with unlimited power

To vary the offspring in character, by mutual

Inexhaustible interchange of the transmitted genes¹.

On this basis the chance re-combinations of a definite number of unalterable factors will yield, as a consequence of sexual reproduction, a wide variety of patterns in the ensuing generation, as dissimilar as the figures formed by shaking the coloured chips in a child's kaleidoscope.

As one of us has argued elsewhere, "probably the most convincing proof of the genetic determination of abilities is the appearance of extremely dull children in the families of the well-to-do professional classes and of extremely bright children in families where both the cultural and the economic conditions of the home would, on the environmentalist's hypothesis, condemn every child to manifest failure". With the Mendelian hypothesis this is just what we should expect. An important corollary follows. On this hypothesis a far

¹ Robert Bridges, *The Testament of Beauty*, iii, 172-4.

greater proportion of the child's inherent characteristics are transmitted to him through his parents than we should infer simply from the resemblances between them as measured by coefficients of correlation.

There is a further reason why a preoccupation with mere resemblances may easily be misleading. The family is the chief instrument for transmitting, not only innate biological characteristics, but also those intellectual, moral, and social characteristics that are acquired afresh by each generation. A few of our critics expressly recognize this fact, and even turn it against us. But here their arguments merely echo what we ourselves have always stressed. It was a point emphasized in the very first of our contributions ([2], p. 172); and we regarded it as a conclusive reason for not accepting test-results at their face value. The older types of test were particularly open to such influences. In the earliest studies of the Binet tests, for instance, it was noted that "tests which depend on a wide vocabulary or on items of skill and information imparted during early life in a cultured home" are apt to favour the child of intellectual parents regardless of his innate capacity; on the other hand, certain tests of a performance type appeared, if anything, to favour the child from the non-intellectual type of home ([4], pp. 194f., and [6], p. 59).

It was largely for this reason that, in formulating our theory, we have always attached special weight to the distinction between genotype and phenotype—a distinction which practically all our critics ignore, although it is basic to the whole Mendelian theory. There are, in fact, no such things as hereditary *characters*; there are only hereditary capacities or tendencies: and the modern geneticist continually insists that "almost every character can, and in fact nearly always does, show heritable and non-heritable variation simultaneously" [12]. To define 'intelligence' as 'what intelligence tests measure', and then argue that "the family background and the environmental influences that result are more than sufficient to account for all the observed differences in a mental character like intelligence"—as Dr. Woolf, Dr. Heim, Dr. Fleming, and several other critics seem to do—is to miss the main point of our argument. No doubt, a few human characteristics, such as eye-colour or serological type, are virtually unaffected by environmental conditions; and these, as we have suggested, might profitably be used as 'markers' to determine in fuller detail the transmission of mental tendencies. But no traits in which the psychologist is interested are likely to be determined either exclusively by inheritance or exclusively by environment; and for much the same reason the mere fact that a particular trait is demonstrably affected by environmental and post-natal conditions is no proof that it may not also be profoundly influenced by heredity.

The wide-spread notion that the evidence for mental heredity rests almost entirely on the study of family resemblances seems largely due to the work of Karl Pearson. Pearson, unlike Galton, strongly opposed what Galton had called the 'particulate theory', and clung firmly to the traditional doctrine of 'blended inheritance'. On this hypothesis it was natural to infer that the coefficient of correlation, as an index of resemblance, supplied the best measure

of heredity. Even today most psychological and educational writers who discuss the subject still implicitly accept the Pearsonian conception; and few seem to realize that any other interpretation of heredity is available. As a result they commonly take for granted that anyone who employs Pearson's statistical procedures, and calculates coefficients of correlation, must necessarily subscribe to Pearson's hypotheses and conclusions. Dr. Fleming [16], for example, describes "Pearson, Burt, Terman, and other workers in the first few decades of the twentieth century" as "believing in the overriding importance of hereditary differences", and apparently supposes that all of them relied on the same type of evidence and adopted the same preconceptions. Dr. Lewis [24], in criticizing the application of a multifactor theory to such traits as intelligence, similarly assumes that its advocates start with the postulates laid down by Pearson. Both critics have evidently overlooked the objections which Burt [2] urged against Pearson's sweeping generalizations, and the arguments which he adduced in favour of the Mendelian doctrine, which Pearson with equal emphasis repudiated.

The Need for Specific Investigations on Specific Characteristics

There is another cautionary principle to be learnt from Mendel. Pre-Mendelian discussions of heredity nearly always dealt with organisms as wholes. When a hybrid animal was bred from two different parental strains, earlier biologists (including Darwin himself) were content to note the general similarity between the different creatures, much as they did in deciding whether a newly discovered animal or plant should be assigned to this or that species. Mendel, on the other hand, in his work on hybridization, announced at the very outset that it is essential to deal separately with separate characters. Psychology has been slow to adopt the same principle. In their references to the genetic differences between the sexes, for example, psychologists, educationists, and sociologists, even in the earlier decades of the twentieth century, still argued the question in the most comprehensive terms: do women display an all-round inferiority to men, or are they the equals of the male? The notion that the average woman might be genetically inferior in some capacities, superior in others, and perhaps qualitatively different in yet other respects, or that certain characteristics might be definitely sex-linked, has been confined chiefly to what Dr. Maberly calls the 'psychometric school' [3]. Most medical psychologists, indeed, sharply contest the 'atomistic view' on the ground that "each personality, whether male or female, must be viewed as a unique and unanalysable whole, not as a mosaic of particles or factors".

A number of our critics, however, somewhat inconsistently assume that 'the hereditarian school' shares this monistic attitude. The psychometrist, it is said, seeks to assign 'a single overall proportion' which will hold good of the entire range of human activity; and his 'belief in the overriding importance of hereditary differences' is depicted as applying to every form of behaviour in an equal degree. But this version of our views is quite unwarranted by

anything that we ourselves have said. Probably the notion that "Spearman, Burt, and their co-workers use the data of their tests to deduce a single all-embracing estimate for the relative importance of nature and nurture" springs from a misinterpretation of the phrase 'general intelligence'. Spearman, it is true, in his earlier correlational studies maintained that his 'hierarchical tables' demonstrated the presence of a single general factor only—"the sole determinant of an individual's potential achievements in every walk of life", and held that specific abilities, or what were described as such, were simply the effects of acquired knowledge, interest, or skill. We ourselves would wholly agree with those who reject any such simple assumption. Nevertheless, as numerous studies of backwardness indicate, an upper limit for each individual child is undoubtedly set by this innate factor of intelligence, though it is a limit which few children, bright or dull, ever actually attain. Moreover, both limits and possibilities depend partly on special abilities; and, like intelligence, these abilities are also to a large extent genetically determined. Consequently, what is actually attainable in the different subjects of the curriculum will be limited in very different degrees.

Dr. Heim and Dr. Harrison apparently imagine that, when we concluded that "about 75 per cent of the variance shown by tests of intelligence is attributable to innate or inheritable causes", the figure was meant to specify "the proportional contribution of heredity to the individual's general educational and social efficiency" in *every* direction. To maintain any such wholesale conclusion would, we readily admit, be "to maintain what is at once meaningless and absurd" [21]. Indeed the absurdity becomes obvious as soon as we consider what such an inference would entail. Our figure of 75 per cent related to the variance contributed by the innate factor to the results of intelligence tests. We should therefore have to suppose that the 'saturation' or 'loading' for this innate factor in every school subject—reading, composition, arithmetic, manual work, and the rest—would not only be the same, but would always be unity. Now in various investigations on tests of educational attainments we have clearly shown that the saturation coefficient varies widely from one subject to another, and that it varies almost as widely for the same subject at different ages.

Spearman's view no doubt implied, as he himself pointed out, that, in theory, every individual could be ranked in a single series which would remain unchanged for all forms of cognitive activity. And as a result those who advocate the use of intelligence tests in the school have generally been credited with much the same doctrine. Not unnaturally, the notion that each human being is stamped at birth with his own serial number, and has his place irrevocably fixed in the queue for relative efficiency, is repugnant to most social writers of the present day, whose sympathies lie usually with those in the lower ranks. However, as must be clear from our previous publications, we should be the first to reject this modernized doctrine of predestination. Human beings, if they are to be classified at all, will have to be arranged, like the abilities they

exhibit, in a ramifying hierarchy, not in a one-dimensional scale: and if some particular individual is weak in one type of ability, there are others in which he may quite possibly excel. Thus our reply to the problem raised by Heim and Harrison would be that very succinctly expressed by Haldane: "the question of the relative importance of nature and nurture has no general answer, but only a very large number of particular answers" ([18], p. 36).

The Analysis of Assessments of Intelligence and the Analysis of Tests

However, in this paper, as in the preceding, we shall be concerned, not with special abilities, but only with general intelligence. Here once again it is to be remembered that what the psychologist can directly observe or measure is only the phenotype, never the genotype. Hence our first aim has been a practical one—to estimate the relative efficiency of the methods available for assessing intelligence. We might almost borrow Galileo's explanation when he repeated Archimedes' experiment: "I am not trying to discover how many ounces of gold there may be in the crown, but simply how much accuracy there is in the method". For this purpose we have endeavoured to ascertain the comparative merits of the two most obvious procedures: first, the use of standardized tests, with the marks taken just as they stand; secondly, checking and correcting the marks so obtained in the light of the available evidence.

Now, what our critics chiefly wish to challenge is the efficacy of the tests in common use. Many are disturbed by the idea that, merely on the basis of a Binet I.Q., some unfortunate child may be permanently labelled 'dull', 'subnormal', or even 'mentally deficient'; others fear that grave injustice may be done to youngsters from the underprivileged classes, if selection for grammar schools is determined mainly by their performances in tests at the 11 plus examination. We ourselves, however, are perfectly willing to admit that, as a means of estimating genotypic differences, even the most carefully constructed tests are highly fallible instruments, and that their verdicts are far less trustworthy than the judgments of the pupil's own teachers; the outstanding merit of the tests is that, unlike the teacher's judgments, they are comparable from one school to another.

Accordingly, our investigations have been directed towards discovering, first, how far such tests are likely to be affected by irrelevant influences, and secondly how far existing methods of assessment can be improved upon by more comprehensive devices. Most psychological investigators have concentrated on the first of these two questions only. Almost inevitably the investigator who enters the schools from outside is obliged to confine himself to short and simple procedures—group tests of the usual type. And here lies the unique advantage of an officially appointed psychologist who is himself a member of the education authority's staff: he not only knows at first hand what are the conditions affecting the examinees in the various schools, but also possesses the authority to extend or repeat his tests and his interviews, and to require from teachers the detailed information and the further assistance that he needs.

The data we have used for the purpose of our various analyses have been secured in this way; and, having satisfied ourselves that by these means we can reduce the disturbing effects of environment to relatively slight proportions, we have gone on to inquire whether the data so obtained—the frequency curves, the bivariate distributions, the correlations between relatives brought up under various conditions (twins, cousins, children and their parents, siblings living in the same home, in different homes, in residential institutions, and the like)—are consistent with the hypothesis of multifactor inheritance. That has formed our chief theoretical problem.

Intelligence and Social Class

Most of the critics we have cited appear to confuse these various issues. This is especially noticeable in discussions on the relations between intelligence and social class. The hereditarian holds that, in communities where there is a good deal of freedom to move from one socio-economic group to another, "the class of an individual depends largely, though by no means wholly, on his innate characteristics and on those of the family from which he springs" [27]. The environmentalist retorts that the class of an individual is "the main factor determining his performance in those tests of intelligence by which his innate characteristics are supposedly assessed". After all, it is said, intelligence tests are for the most part designed by academic psychologists, that is, by persons with the cultured outlook of the 'privileged classes'. As a result, they have "a strong bookish bias, and are decidedly 'U'" (i.e. upper class) "in their tone and vocabulary"; and to assume that such tests "measure intelligence and that what intelligence tests measure is innate" is (so Dr. Heim declares) to be guilty, not only of 'verbal imprecision' but also of 'personal bias' [21]. Professor Zangwill has recently come to the defence of the test-constructor, and suggested that his bookish bias may be justified by the fact that our present civilization itself is built up on books: in the near future, he thinks, a technological age may lead the psychologist to prefer a more practical brand of test. But the psychologist selects his tests and his test-items, not in accordance with his subjective notions, but on the basis of an objective item analysis and of the correlations of his tests with the main underlying types of 'cognitive' (directive) process—perception, association, coordination, abstraction, reasoning, noetic synthesis, and the like. Unfortunately in actual practice the most effective modes of testing cannot always be used; and the defects of the tests adopted for workaday purposes result usually from the unavoidable need to adopt something that entails no great expenditure of either time or money.

All this, however, Dr. Heim seems to overlook. To support her views she cites the figures published in the inquiry on 'Ability and Income' by Burt [7], which, she says, were "acclaimed as proof of the supremacy of heredity and the rightness of the existing order of things For such people" (the psychometrists of the hereditarian school) "equality of opportunity consists

in affording individual children the same opportunities as their parents"; and thus the investigator mistakenly infers that "'the haves' inherit from 'the haves'" their apparent intelligence, whereas in point of fact all that they inherit are "more books in the home" and "the means to pay for better schooling". The figures recently collected by Mrs. Floud [17] and Mr. Flann Campbell [10] in regard to the social composition of the child population in secondary 'grammar' schools have been interpreted in much the same way.

Now we cannot help feeling that it is Dr. Heim who is really guilty of "verbal imprecision and personal bias". When the psychometrist states that this or that group test 'measures intelligence', he merely means that it yields a more probable estimate of each individual's intelligence than could be obtained for the same group with the same expenditure of time and money. Dr. Heim interprets the phrase as though it meant 'this test measures intelligence *and nothing else*'. None of those she criticizes have ever assumed that "what intelligence tests measure is innate". And even the most hasty perusal of the article that she quotes will show that Burt and his co-workers were far from claiming that the figures there given supported "the rightness of the existing order of things". Indeed, their final conclusion was that, among the non-fee-paying population, "about 40 per cent of those whose innate abilities are of university standard *fail* to reach a university, whereas an equal number from the fee-paying classes receive a university education to which their innate abilities alone would scarcely entitle them" (*loc. cit.*, p. 98). Nor is it true to suggest that "the psychologists who design and recommend such tests are themselves members of the privileged classes". In point of fact, as they have gratefully recognized, those who have been most active in this field would themselves have been unable to enjoy the benefits of a public school or a University had it not been for the scholarship system; to quote Godfrey Thomson, this fact was for him (as it was for several others) the main "incentive . . . in endeavouring to improve methods of selecting pupils" ([37], p. xv). A glance at the reports of the L.C.C. Psychologist will show that in actual fact such tests have served far more frequently to save the backward and the merely retarded from being certified as mentally deficient and to award the bright pupil from a poor home a free place in a 'secondary' or grammar school when his performance in the older type of scholarship examination would probably have excluded him.

But in any case much of this criticism is really beside the point. No competent psychometrist would nowadays rely on results collected from tests applied primarily for practical purposes, in order to solve his own theoretical problems. At best, in an inquiry into the extent or nature of mental inheritance, the figures supplied by official examinations can furnish nothing but confirmatory evidence. And the material which we ourselves have chiefly used was obtained from surveys planned with the theoretical issues specifically in view.

Some justification for the protests made by Dr. Heim and other environmentalists can no doubt be found in the exaggerated claims of the older eugenis-

vehement assertions on the one side inevitably provoke vehement dissent on the other. Pearson, for example, more than once declared, on decidedly tenuous evidence, that "the influence of environment is nowhere more than one-fifth of that of heredity, and quite possibly not one-tenth of it" [31]. *Sed latet dolus in generalibus*: and this sweeping generalization trapped him into practical corollaries which not unnaturally tended to discredit the whole movement: "a heredity-hunt", as G. K. Chesterton observed, "would be a greater menace than a heresy-hunt" [11]. A few years later Dampier Whetham, one of the most active champions of eugenics, stoutly opposed all 'competitive examinations' on biological principles; "scholarships" (he warned his readers) "have their dangers when used to raise those who win them out of their natural class; they were good sociologists, as well as good divines, who taught the child to 'learn and labour truly, and do his duty in that state of life into which it has pleased God to call him'" [40]. "Francis Galton himself", Professor Hogben reminds us, "was a man with a strongly aristocratic bias". And perhaps it was not surprising if "eugenics thus became identified with a system of ingenious excuses for combating the amelioration of working class conditions" [22]. Yet, when all is said, it seems a little unfair to saddle psychologists of the present day with the exaggerated views and social prejudices of the pioneers of fifty years ago. Scientific issues cannot be solved by turning them into a kind of erudite class-war—patricians *versus* plebs. After all, the basic error committed by the earlier hereditarians was much the same as those of their present day critics. They equated 'observable intelligence' with 'inherited intelligence'—the phenotype with the genotype; and, if their theory of inheritance was grossly at fault, that at least was pardonable when genetics was still in its infancy.

III. METHODOLOGICAL CRITICISMS

General Criticisms

The main methodological argument may be briefly summed up as follows. What the psychometrist calls 'genotypic intelligence' is avowedly an abstract and hypothetical 'factor'. Now, hypothetical abstractions, like 'factors', cannot be empirically measured; even if they could, the 'factors' of the statistical psychologist are admittedly non-existent figments; and in any case all endeavours to establish the existence of an innate factor of intelligence have so far completely failed; therefore, any further attempt to investigate or measure it is plainly doomed to failure; and the quest may as well be abandoned.

Let us take these allegations one by one.

The Alleged Impossibility of Measuring Hypothetical Quantities

No one denies that strictly speaking "an innate quality, like intelligence, from its very nature cannot itself be measured". But in that case, says the critic, "there can be little point in discussing a hypothetical innate ability unless we can measure it", particularly as the only method of verification available must itself depend on the assumption that the quantities involved are

to be measured. This difficulty was urged by several contributors to the Symposium arranged by Section J (Psychology) of the British Association at its Newcastle meeting; and more recently Dr. Maddox and two or three of our correspondents have maintained [26] that any quantity which cannot be directly measured or 'operationally defined' must be 'metaphysical', and that it is "a postulate of modern scientific methodology that all such unverifiable concepts should be discarded out of hand, together with the ether and similar hypothetical figments".

Now we doubt whether the majority of scientists of the present day would subscribe to the 'postulate' on which this argument is founded, although twenty years ago the proposal to dispense with all but directly verifiable concepts enjoyed for a time a widespread vogue. Contemporary physics, particularly quantum physics, is full of hypothetical quantities which at best can only be estimated with varying degrees of accuracy. Energy, to take the most familiar concept, cannot itself be directly measured. Like intelligence, it is, by definition, a 'capacity', namely, a 'capacity for work'. Consider the ordinary methods of determining the amount of energy supplied, say, by a Leclanché or Daniell's cell. As the student of physics is warned, the laboratory methods commonly employed are valid only on the assumption that the energy from the chemical changes is wholly converted into energy of current. But, strictly speaking, this is never the case; and various 'errors of measurement' have always to be allowed for. This is exactly parallel to the psychologist's problem. The 'capacities' with which he is concerned can never be directly measured, but only assessed. Nevertheless, provided he can devise some means of judging the probable margin of error, this in no way precludes him from using such assessments in his statistical proofs.

The Alleged Non-existence of Mental Factors

Many of our methodological critics would readily allow that certain hypothetical quantities may, in some sense or other, be measurable; but they insist that such quantities form a strictly limited group. Even if there is no direct observational evidence that the quantity in question *does* exist, yet (so they argue) that quantity must be a thing that *might* exist; and, when it is introduced as part of a working theory in science, there ought in addition to be some factual evidence to show that it probably *does* exist. Critics who pursue this line usually quote Thomson's warning about the dangers of 'reifying' abstract factors: "the psychologist's factors are statistical artefacts; a man does not have a *g* in the sense that he has a liver" [37]. Moreover, as they point out, the variable that appears in our tables is "an abstract factor twice removed": it is the genetic component (or sub-component) of the general cognitive component. Such a rarefied abstraction, they contend, is "nothing but the ghost of a ghost", and can have "no real existence".

Here it seems helpful to distinguish between the mathematical model on which our proofs are based and the 'real' or concrete phenomena to which the

model refers. Practically all scientific theories are founded on symbolic models of this kind—highly simplified constructs formed out of highly abstract concepts, skeletons with the flesh all scrupulously removed. They are essential for the preliminary deductions; and, in order to apply or verify the corollaries so deduced, all that is logically necessary is that the algebraic variables defined in the model should correspond with, or, as the phrase goes, 'be isomorphic with', the essential features of the real phenomena. There is no need for them to be directly correlated with what is actually observed or measured: they may simply represent what Broad has called 'dispositional properties' ([1], pp. 434f.)¹. And, as he reminds us, it is distinctive of modern, in contrast to mediaeval science, to interpret causal properties of this kind (including what mediaeval writers would have called the 'powers', 'faculties', or 'capacities' of the mind) in terms of the minute spatio-temporal structure of the material substances that presumably serve as vehicles for such properties. That, as we have argued elsewhere, would be a perfectly feasible way of interpreting the particular 'dispositional property' which we have designated 'intelligence'.

But, of course, this version of the situation demands a separate set of proofs, to show how the factors in the model are related to, or represent, the characteristics in the actual persons that we test. Oblivious of this gap in their arguments, statistical psychologists are too apt to confine themselves to purely statistical considerations; and are often content to identify their 'factors' straightaway with supposed psychological characteristics on no other evidence than that afforded by the names of the tests employed. In our view, such identifications always call for the support of non-statistical evidence—observational, introspective, experimental, neurological, or the like [8].

In an instructive paper, Mr. Miles has argued that hypothetical entities like the factor of intelligence may be described as having 'real existence' "only when they are not 'factors of the mind', but 'factors of the body', i.e. dependent on the 'discoverable behaviour of genes or neurones'" [29]. We fully agree that this is by far the most satisfying way of conceiving the dispositional properties that concern us here; and we hold that a plausible interpretation of this kind is actually furnished by the multifactor theory. Could this theory be adequately verified, we might even define an individual's innate intelligence (γ) as proportional to the number of γ -genes in his genetic constitution, where a γ -gene is conceived as part of the ultra-microscopic structure of the chromosome, and may itself be defined as an

¹ For our purposes it will be sufficient to define a 'dispositional property' as one which requires a hypothetical proposition to describe it. Thus, when we say that a steel spring has the property of 'elasticity', we mean: "if such a spring is pressed, pulled, or strained, and then released, it will (unlike a piece of putty) return more or less to its original shape". The phrase 'more or less' implies that the property has varying degrees; and therefore, although by definition a hypothetical abstraction, is susceptible of measurement; in fact, the degree of elasticity may be expressed by various ratios or 'moduli': e.g.

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\text{pull per area of cross section}}{\text{change in length per unit length}}$$

allelomorph whose presence adds a small positive quantity to one and the same specifiable mental capacity, namely, the capacity to perform all kinds of cognitive task. It will no doubt be objected that to count the actual number of such genes would be an impossible feat. True; but it is equally impracticable to count the number of molecules in one cubic millimetre of a given substance; yet the molecular density of a gas is defined in terms of the number of molecules contained in such a volume. In both cases what the scientist does is to make, not a direct count, but an indirect estimate of the number in question.

The Alleged Non-existence of 'Innate Intelligence'

Dr. Maddox [26] endeavours to reinforce the criticisms we have been discussing by giving the methodological argument a more empirical turn. After prolonged research by various methods, he says, "investigators have failed to produce any convincing evidence about the genetic components of intelligence". Hence our own "attempt to resuscitate the nature-nurture controversy" he condemns as quite "anachronistic". A wiser methodology would leave such purely conjectural factors in the nebulous limbo to which they belong, and would concentrate exclusively on environmental factors. "Environmental factors", he points out, "are capable of control; and for this reason are of greater interest to educators". He therefore fully applauds the conclusion expressed by H. E. Jones "that potential contributors should now be advised that the nature-nurture controversy has been shown to be an unproductive field of research".

This would be an easy way of cutting the Gordian knot. But a policy of contented nescience would mean renouncing every hope of solving numerous problems which are of first importance alike for theory and for practice. The mere fact that environmental conditions are more amenable to control, and hence of paramount interest to the educator, cannot of itself absolve the psychologist from the task of investigating the possible influence of genetic factors. In dealing with any particular child¹, it is as necessary to consider what cannot be done as to know what can. If Tommy's backwardness is due to an innate lack of intelligence, then it is useless, and worse than useless, to strive to transform him into a normal pupil. If Harry is not endowed with the ability requisite for the work in a grammar school, it is folly to transfer him thither simply because there is a slender chance that "even the dullest youngster may be a genius in disguise". It is only in fiction that the hundred-to-one outsiders romp home in the final lap, and leave the experts gaping.

Our own explanation for the failure of previous research would be very different from that which Dr. Maddox has offered. The obvious reason surely is that the investigators to whom he refers were all thinking in terms of out-of-date concepts of heredity: even had they wished to substitute a Mendelian

¹ For conflicting opinions on this point, see the articles and correspondence on 'Dullness and its Causes' in the recent issues of the Times Educational Supplement, October 19, 26, November 3, 10, 1956.

theory, they possessed no adequate techniques for coping with multifactorial problems. Dr. Maddox, however, apparently believes that, from the very nature of the issues involved, it is impossible to secure evidence that might justify a hereditarian interpretation; and his scepticism, as he points out, can claim the sanction of most American psychologists. Nevertheless, it no longer commands the support of American geneticists. Professors Goldschmidt, Snyder, Calvin Hall, and Calvin Stone all deplore the neglect of genetic investigation by psychologists of the present day. Goldschmidt himself traces this general lack of interest to the fact that empirical psychology began as an offshoot of philosophy, at a time when associationism held the field: translated into physiological terms, the associationist theory gave rise to a 'doctrine of universal conditioning', which has lately dominated the American behaviourist school; and as a result practically all individual differences are interpreted in terms of postnatal experience. "This", he writes, "may have been responsible for the extreme belief of our contemporary psychologists in the power of environment". Nevertheless, "with a little prodding from genetics, both human and animal psychology might increasingly assimilate genetic notions, and be stimulated by them for mutual benefit"¹.

What, then, in the most general terms, are the methods of the present-day geneticist, and what grounds have we for supposing that they cannot be profitably borrowed by the psychologist? Broadly speaking, there are four main lines of approach: (i) cytological analysis—the study of cell structure; (ii) the method of uniform environment with varied heredity; (iii) the method of uniform heredity with varied environment; (iv) genetic analysis—the study of the effects of the segregation and recombination of genes in successive generations (cf. [36]). Man is a slow-breeding animal of low fertility; he cannot be readily subjected to experimental control; and his behaviour is governed by post-natal experience to a larger extent than that of any other creature. These peculiarities render the psychogeneticist's task more complex; but they do not make it impossible. The validity of such methods for the study of mental characteristics is a matter to be decided, not by the abstract canons of some preconceived methodology, but by the pragmatic principle of giving them a trial, and noting whether they yield any worth-while results.

The evidence we have offered in support of our own hypothesis has been derived from all four lines of approach. (i) In the case of man, cytological knowledge is still lamentably defective; yet it is sufficient to demonstrate that the mechanisms of heredity and variation are essentially the same in the human species as in all other bisexual creatures. We have relied on it chiefly to show

¹ *Genetics in the Twentieth Century* (1951), p. 7. In the same volume Professor Snyder recalls how, "after a short flurry of interest in human heredity during the first decade or two of the century, that interest, particularly in America, proved to be short lived" (*loc. cit.*, p. 370). Quite recently, however, centres of research in human genetics have been established in the United States, and by 1948 an American Society of Human Genetics had been formed. Although hitherto engrossed with the problems of physical and pathological heredity, its members are "already beginning to ask whether it may not be time to embark on genetic studies of normal mental characteristics".

(a) that there is a strong antecedent probability that a basic character like intelligence, which appears to be continuously graded and normally distributed, is determined, or at any rate influenced by, a multiplicity of genetic factors and (b) that the number of genes or 'factors' is large enough to provide the wide range of differences and the smooth distribution curves that are actually observed. (ii) Studies of children brought up in residential institutions where environmental conditions are relatively uniform have disclosed wide individual differences in intelligence that cannot possibly be ascribed wholly to environmental influences, and are closely correlated with those exhibited by the parents. (iii) Studies of monozygotic twins reared in different environments have been used to assess the effects of environmental differences on (a) intelligence tests of the ordinary type and (b) adjusted assessments based on all the information available for each individual child. (iv) In the main, however, the procedure adopted by the human geneticist must necessarily be statistical; and here the methods of genetic analysis developed by the modern biometric school have already proved their efficiency in analogous fields of work. Today, geneticists who work with cattle or with crops are constantly attempting to estimate the strength of whatever genetic quality it may be that makes for a good yield in this or that breed and in this or that strain: there can surely be no *a priori* objections to extending their methods, with appropriate modifications, to the problems of human psychology.

IV. STATISTICAL TECHNIQUES

General Nature of the Techniques Employed

This brings us to the type of criticism most frequently advanced, namely, that directed against the biometric techniques that we ourselves have applied. During the earlier part of the century the statistical treatment of genetic problems, initiated by Galton, Pearson, and Weldon, was quickly overshadowed by the more direct and fruitful techniques introduced by the experimental and cytological investigators of the Mendelian school—particularly Bateson and his followers in this country and Morgan and his fellow workers in the United States. The interest in statistical research began to languish; and its neglect has been widely interpreted by psychologists and others as a sign of sterility. Today, however, practically all geneticists are fully agreed that, in attacking problems of multifactorial inheritance, statistical techniques are indispensable; and those developed by Fisher and his followers now command almost universal assent.

Yet they have not escaped all criticism, even at the hands of biologists themselves. In this country the most prominent critics have been Hogben and Woolf, who have paid special attention to the study of human intelligence. For an authoritative examination of their views the reader must be referred to the replies of the writers they have criticized. But a brief glance at the difficulties which they emphasize seems necessary here.

A. *Objections to the Analysis of Correlated Variables.* Several of our critics cite the final chapter of Hogben's book on *Nature and Nurture* [23] as if it completely disposed of any attempt to apply Fisher's methods to such problems as our own. Dr. Maddox, for example, in discussing what he describes as our 'balance sheet for nature and nurture', claims that Hogben's arguments conclusively demonstrate that the whole procedure is 'logically untenable'. Hogben himself never goes so far as this. Indeed, in the chapter quoted, he expressly admits that "the technique of correlation can be used to draw attention to the existence of genetic differences or of differences due to environment", and declares that "the importance of Fisher's hypothesis of cumulative factors is beyond dispute". His chief contention is that Fisher's method of analysis, at any rate in its original form, is legitimate only when we are able to "deal with nature and nurture as independent variables"; when nature (or 'heredity') and nurture (or 'environment') are themselves correlated, then (so he maintains) the results may prove deceptive. "The data on which Fisher's own analysis for stature is based were", he points out, "collected from a comparatively homogeneous social group", i.e. from a group in which the relevant environmental conditions were much the same for all; but were we to apply it to "human populations in general", we should be forced to recognize that the wide range of environmental differences might falsify any values "theoretically deduced on the assumption that the environment is homogeneous". Hence, although such methods may safely be trusted to indicate the *presence* of genetic factors, we must be cautious about using them to indicate its *amount*; and thus, whatever significance Fisher's attempt to draw up "a balance sheet of nature and nurture" may have for the study of human characteristics, "it cannot entitle us to set limits to the changes that might be brought about by regulating the environment"¹.

In his larger volume on *Genetic Principles* Hogben sounds a more optimistic note. He assures us that "from a theoretical standpoint the prospect is brighter" than it formerly appeared. In his opinion, it is now both possible and necessary "to develop methods which will enable us to assess the relative importance of nature and nurture in a specified range of conditions". Among the procedures available he includes the study of "observed differences in performance of psychological tests", such as he himself has used in his investigations on twins; and he advocates five particular types of inquiry "which, when applied with proper safeguards and interpreted with discrimination, will", he believes, "yield unequivocal data". These are the study (i) of twins and siblings, (ii) of adopted children, and of the effects of (iii) inbreeding,

¹ The book quoted by Dr. Maddox and Dr. Heim consisted of semi-popular lectures, addressed to a medical audience nearly twenty-five years ago, and was concerned, not so much with the technical merits of Fisher's mathematical procedure, but rather with the practical inferences which had rashly been drawn from it. Hogben's own views are more systematically presented in the larger textbook cited in the next paragraph.

It should be added that, since the above was written, Dr. Maddox has slightly revised the wording of his argument: but his essential conclusions remain the same. Our quotations were taken (with his permission) from the earlier version.

(iv) outbreeding, and (v) order of birth; and under each heading he cites one or two model researches, usually from the domain of animal genetics. These lines of inquiry, it will be seen, are much the same as we ourselves had tentatively employed in our earlier investigations. And, though Hogben himself might not approve the details of our work, there is no difference of principle between us. In short, the prophet whom Dr. Maddox has called in to curse such enterprises seems more disposed to grant them his blessing.

Hogben ends with a plea which we ourselves should like to second as strongly as we can. After deploring the lack of "University machinery" for giving instruction in modern genetics to students of medicine, social psychology, and allied subjects, he goes on to urge, in the most convincing terms, the importance of independent provision and special centres for research in the field of human genetics. It is, as he says, a subject which demands "organized teamwork on a very large scale", including the "collaboration of geneticists, statisticians, educational psychologists", and other qualified persons. It will therefore be costly. Yet "what we know already liberally encourages us to hope we may eventually know much more".

All this seems far more reassuring in its tone than Dr. Maddox's quotations might suggest. No psychologist would quarrel with the cautionary remarks which Hogben makes in criticizing the unwarranted deductions drawn by the older enthusiasts. We ourselves have never sought to apply the method to "human populations in general". And the difficulties that arise when environment is allowed to vary unquestionably call for special treatment. They can, we believe, be satisfactorily met in two alternative ways: (i) by modifying the design of the experiment; (ii) by amplifying the statistical technique.

(i) When planning the inquiry, we may seek to exclude any irrelevant correlation with environment by selecting either the sample of persons or the sample of tests. Thus, when using tests that are not 'culture-free', we could for purposes of theoretical research select children coming from a restricted type of environment in the hope of eliminating in advance any cultural differences that might affect the tests. On the whole, however, we think it best to substitute, where possible, tests (or other methods) of such a nature that cultural differences are not likely to affect the results.

(ii) In studying, for more practical purposes, the value of marks obtained with tests of the ordinary type, we have endeavoured to supplement Fisher's original procedure so as to cope with the further variables involved. To check the adequacy of the method, we proposed an alternative analysis in terms of the so-called parallelogram law. Here, instead of starting with the correlation between assessments and environmental differences, we deduced it; and we showed that in each case the coefficient so obtained agreed, as closely as could be expected, with the empirical value computed directly from ad hoc ratings.

Dr. Maddox himself raises one or two further objections to Fisher's procedure. Since, he says, when judged by the correlation actually obtained,

ancestral heredity will account for little more than 50 per cent of the total variance in bodily characteristics, Fisher is obliged to "invoke other factors, such as segregation and assortative mating, to avoid attributing the balance to environmental factors", and as a result has grossly underestimated the effects of favourable and unfavourable environment and the probable interaction between the environmental and the genetic components. Moreover, so Dr. Maddox contends, even the "initial step"—namely, ascribing the correlation between children and their parents to genetic factors—"is suspect, because of the correlation of heredity with environment".

To these further arguments there are three simple replies. (i) Segregation and assortative mating were not arbitrarily invoked. Assortative mating is an ascertained fact: and without segregation the hypothesis would not be Mendelian, but pre-Mendelian. Does Dr. Maddox really hold that Fisher should have discarded the concept of 'segregating factors' in favour of the older theory of 'blending factors'—a theory rejected by every modern geneticist? (ii) The data for bodily characteristics which Fisher used were obtained, as we have seen, from a "fairly homogeneous middle class group". Hence with this particular group the supposed differences between "favourable and unfavourable environment" did not in fact exist. Further, it is quite erroneous to assume that the effect of environment inevitably increases the correlation found. As Yule points out, both environment and dominance may tend to reduce it. (iii) Finally, Fisher readily admits that some of the genetic components are only "imperfectly additive", and has therefore introduced a specific procedure which enables him to treat the irregular effects of interaction as an independent residual component.

B. Objections to Second Degree Statistics. The most direct attack on the type of approach that we have used is contained in a provocative paper contributed by Dr. Bernard Woolf to the Edinburgh Colloquium on 'Quantitative Inheritance' ([33], pp. 81-102). His argument is divided into two parts: in the first he challenges the principles and assumptions adopted by nearly every British writer who has discussed quantitative inheritance in man—Fisher, Fraser Roberts, Thomson, and Burt; in the second he examines the specialized techniques initiated by Fisher, developed by Mather, and "common to current workers in biometrical genetics". The final outcome, so Dr. Woolf tells us, is to enforce two main contentions: "the first is that real genetic situations with environmental and other interactions include too many variables to be amenable to the kind of analysis developed by Fisher and his school; the second is the utter inadequacy of orthodox analysis of variance, when confronted with a real world not arranged in Latin squares with all effects strictly additive".

(i) *The Inheritance of Intelligence.* All too often our various critics seem to take their notions of what their opponents have said, not from their opponents' publications, but rather from each other; and Dr. Woolf's somewhat garbled version of our views has been so readily accepted by later writers that it is necessary at the very outset to emphasize how often he misinterprets what we

have actually said. He opens with a detailed discussion of an ever-popular problem—the alleged decline of national intelligence; and attacks what he takes to be the solution propounded by those who adopt the ‘multiple or additive factor hypothesis’. Nowhere does he explicitly state how he himself understands the term ‘intelligence’; but his comments imply that he sides with “those who prefer to re-define intelligence as ‘what intelligence tests measure’”. However, little or no interest can attach to the question whether the mere capacity to answer tests has improved or diminished. The issue raised by the writers he quotes, and revived by the inquiries of the Royal Commission on Population, related to what the Commission itself called ‘innate intelligence’. Burt’s abridged memorandum, to which Dr. Woolf chiefly refers, gives a formal definition in the opening paragraph; and the Appendix ‘On the Meaning of Intelligence’ emphatically rejects the alternative, which Dr. Woolf seems to prefer ([41], pp. 68–75). Of course, if Dr. Woolf insists on starting off with a ‘re-definition’ of the crucial term, it becomes the easiest thing in the world to show that our argument leads to a *reductio ad absurdum*.

However, besides misinterpreting the key-word in the discussion, he also gives an unfair twist both to the evidence adduced and to the final conclusion. In summarizing the evidence on which his opponents chiefly relied, he tells the reader that “all their predictions have been *purely inferential*, based on *test scores* taken at *only one point* in time” (*loc. cit.*, p. 84, our italics). This statement contains at least three misleading implications. First of all, in the London investigations at any rate, ‘the predictions’ were not based solely on ‘test scores’. In the memorandum and elsewhere a clear distinction was drawn between ‘the crude test results’ and the ‘adjusted’ or ‘final’ assessments. In doubtful cases different types of test were used at different ‘points of time’; and in every case the assessments were submitted to the teachers, and, where necessary, corrected in the light of fuller information.

Nor is it true that ‘the predictions’ were ‘purely inferential’. If Dr. Woolf will turn to the section in the Memorandum entitled ‘Estimating the Amount of Apparent Decline’, he will see that *two* methods of prediction are discussed. The first, headed ‘Indirect Estimation’ (p. 69), *was* ‘purely inferential’; but the calculations were immediately qualified by the comment that on such matters armchair deductions cannot be trusted without confirmation. Had the calculated changes actually occurred, they would have involved a marked increase in the number of defectives and an equally marked decrease in the number of scholarship winners; and this, it was argued, “the older head teachers and officials would assuredly have noticed”. An alternative method was therefore used in the hope of “securing direct confirmation”. This is described under the heading ‘Direct Estimation’ (pp. 61f.), and consisted of repeated surveys carried out at two or more widely separated ‘points of time’. The result was a very different set of figures. The conclusion ultimately reached was that “far more extensive studies are required”.

After sounding his 'stern warning' against the 'fallacy' of indirect inference, Dr. Woolf goes on to say: "Still less was it legitimate for Thomson and Burt, *without any tests whatever of the parents*, to visit upon them the marks scored by their children Their anticipations of decline might have been justifiable as pessimistic speculations, though the *great confidence* placed in them would not have been" (p. 86, our italics). But in each of our surveys, assessments were individually obtained for a representative sample of parents, checked, for purposes of standardization, by tests of the usual type ([41], p. 52, [8], p. 172). How else does Dr. Woolf suppose the correlations between children and parents were computed? Moreover, the burden of the whole memorandum was to deprecate any feeling of 'confidence'. And the opening summary plainly announced that the evidence at present available was "far from conclusive". Hence the reiterated plea for research.

In discussing the influence of genetic and non-genetic factors on individual differences in intelligence (or rather on the assessments of those differences), the "first conclusion" that Dr. Woolf wishes to enforce is that "Burt seriously underestimates the amount of population variance attributable to environmental influences". To establish this point, he begins by examining "the study by Barbara Burks on which he (Burt) bases such far-reaching conclusions" (*loc. cit.*, p. 87). Dr. Burks' investigation, we are told, was carried out on only "214 foster children with 105 children as a control group". But the principal basis for Burt's conclusions was not this American inquiry, but a series of investigations conducted with the help of several co-workers, and covering nearly 600 children in "residential schools, orphanages, and foster homes where the inmates are received during early infancy" and including a far larger sample of controls attending ordinary L.C.C. schools. This itself was merely supplementary to a mass of other data—correlations between parents and offspring, siblings, twins, and the like—to which Dr. Woolf nowhere alludes. The American inquiry was only cited at the end of the Appendix to show how an independent investigator in an entirely different country had arrived at much the same figure as that derived from the London surveys.

Dr. Woolf next turns to Burt's table for the mean assessments for children classified according to their father's occupation. He admits that the total range, from those belonging to "higher professional or administrative" classes at one end of the scale to "casual labourers" and "institutional defectives" at the other, is "enormous . . . amounting to more than 30 points". But, he asks, is it not possible to account for "all the observed differences by environmental effects . . . without any need to bring in genetical differences"? To support this suggestion he quotes figures from various sources (including Burt and Thomson); but all that these figures show is that environment may appreciably affect uncorrected test-results—a point which of course was never in dispute. He "freely acknowledges that all these explanations" (i.e. explanations he himself would favour)—"differences in parental attention, birth-rank effect, pre-natal disadvantages, etc.—are speculative"; and certainly the data

he cites would seem rather to indicate that these various 'non-genetic influences', though undoubtedly present, are far smaller than the genetic.

However, this is not our main objection. What we demur to is the tendentious way he frames his question. The implication is that there are only two alternatives: (i) that *all* the differences are to be accounted for by heredity; (ii) that *all* of them are to be accounted for by environment—a most misleading antithesis. Surely Dr. Woolf would not pretend that intelligence differs from every other human characteristic in being totally unaffected by heredity. And certainly he does not deny that differences in the apparent intelligence of children are associated with differences in the genetic constitution of their parents. Why then should we not invert his question, and ask: "is it not possible to account for all the observed differences by the genetical differences without bringing in environmental influences as well?"

But, of course, both questions fallaciously imply that the two modes of causation must be mutually exclusive. For such an assumption there is not the slightest ground. Indeed, it must be obvious to anyone, who is not already strongly biased towards one explanation or the other, that *both* types of influence are generally at work. And the question to be asked is not: *could* the difference be accounted for wholly in this way or wholly in that, but what is *the most probable* way of apportioning the results between the first type of cause, the second type of cause, and the mutual interaction of the two.

Throughout this part of his paper much of the apparent plausibility of the arguments employed rests on a fallacy so often committed by psychologists and others who have not themselves had experience of mental testing. He treats 'the I.Q.' as a kind of absolute measurement which is necessarily the same for the same child regardless of the method by which it has been obtained. Thus conclusions drawn by one writer who is referring to I.Q.'s obtained with group tests are applied to other I.Q.'s obtained by another writer who used individual tests; conclusions drawn from verbal tests are applied to I.Q.'s obtained with non-verbal tests; and conclusions drawn from uncorrected I.Q.'s obtained from any test are assumed to hold good of composite assessments, carefully checked, where the I.Q. forms merely a conventional unit. Hardly ever, when quoting an I.Q., does he mention how it was secured. With the original Binet tests (with which some of the quoted results were obtained) the standard deviation was less than 13 I.Q. points; with the tests used by Cattell it was over 20; for our adjusted assessments (based originally on percentile ranks) we used an arbitrary I.Q. unit with a standard deviation of 15. Dr. Woolf treats these and other I.Q.'s as though they were all equivalent. That, of course, would make nonsense of any re-allocation of the variances.

(ii) *The Analysis of Variance*. In attempting to assess the relative influence of genetic and non-genetic factors, we ourselves relied almost entirely on an analysis of variance; and the second part of Dr. Woolf's paper is devoted to a spirited attack on this procedure.

His first and main objection to what he describes as 'the red herring of variance analysis' is that it commits us almost inevitably to an underestimate

of those environmental influences that he himself has enumerated. It is, he says, a necessary condition of the statistical technique adopted by Fisher and his school that "environment should appear as a random error which can then be eliminated by arithmetical legerdemain rather than as an integral part of the situation to be embodied in the conclusions". Here, beneath the phrase 'rather than' there lurks the same fallacy as before—the tacit substitution of the 'exclusive (or alternative)' *or* for the 'inclusive'. From the standpoint of the statistician, environment may be conceived as influencing individual variations in two distinct ways—first, by its 'systematic' effects (as 'an integral part of the situation'), and secondly by its 'random' effects. These are two *separable factors* (though in fact they are usually found together); they are not two *alternative methods* of treating one and the same factor. Because they are separable, it is possible, by a suitable choice of cases, to get rid of the systematic effects before ever we proceed to the statistical analysis. When this has been done, it is no longer true to say that what is an 'integral part of the situation' now 'appears' as a random factor. It has been eliminated in advance; and the 'random' part is merely what is left. Moreover, in the material with which Fisher was dealing, the elimination could not have been the result of Fisher's statistical technique, since it had already been effected by the manner in which Karl Pearson had actually chosen his sample.

In another part of his paper, Dr. Woolf seems vaguely aware of this point, and puts his criticism in a different form. If, he says, we intend to assume (as Fisher does in the case of bodily measurements and as we have done in the case of intelligence) that the trait to be analysed is "influenced by *both* heredity *and* environment, then environment *must* be given primary place in the analysis: it is not legitimate to postulate genetic differences until it has been shown that environmental influences are unable fully to account for the variations in the trait; even then, heredity can only be invoked to account for the residual differences" (our italics).

Now, had the researches which Dr. Woolf is criticizing been designed to study the various processes by which environment can affect the phenotypes, then the principle he thus prescribes would no doubt have been appropriate. But their purpose was quite different. Fisher's object was to demonstrate that his own theory of human inheritance accounted for the data obtained by Pearson far better than the theory that Pearson himself had championed. We in turn were anxious to see whether much the same form of the multifactor theory might not account for the parallel figures obtained with assessments of intelligence. But we had an ulterior aim as well, namely, to determine, within rough limits, what proportion of the total variance was attributable to genetic factors in the case of (i) raw measurements obtained from tests and (ii) adjusted assessments based on all the evidence accessible for each individual child. Surely it is sound methodology to eliminate so far as possible, and for the rest to randomize, environmental effects when the object is to analyse heredity, just as it would be to eliminate or randomize genetic effects if the object was to analyse environment.

Dr. Woolf, however, declares that "randomization of environmental effects is tantamount to saying that there is no hope of ever understanding that aspect of the situation". Nevertheless, in other researches—e.g. Burt's studies of backward children and of those mistakenly diagnosed as mentally deficient, where the problem was to analyse the manifold ways in which environment might mould the child's mental development and modify his performances in tests of various types—the operation of 'that aspect of the situation' *was* very intelligibly explained. In any case, this part of Dr. Woolf's argument surely flies in the face of his other contention—the contention which he himself places first. He started, it will be remembered, by declaring that "real genetic situations . . . contain too many variables to be amenable to the kind of analysis developed by Fisher and his school". And that is precisely why, on our view, when planning such inquiries, it is essential to simplify the issues, and deal with separate aspects of the situation in separate inquiries. But, if in the same research we try to embody environment 'as an integral part of the conclusion', that must inevitably multiply the number of variables instead of diminishing it.

In order to demonstrate in greater detail the fallacies involved in using 'second order statistics' (i.e., variances and coefficients of correlation) Dr. Woolf proceeds to scrutinize at some length the illustrative example worked out by Mather in his book on *Biometrical Genetics*. This he takes to be typical of the methods adopted by current workers in that subject. He himself contends that in all such investigations 'first order statistics are essential'. Why precisely first order statistics should be exempt from the disadvantages of second order statistics is left rather in the dark. Indeed, as Dr. Mather observes, Dr. Woolf's mode of argument seems to "contribute more to a trenchant style than to a scientific appraisal"; and in his criticisms he too often "actively misrepresents both the assumptions made and the methods used". In regard to the issues with which we ourselves are concerned, there would seem to be little difference between the results obtained by the two types of statistic. Suppose, for example, we wanted to determine whether the test-performances of children reared in foster-homes were superior or inferior to those of children reared by their own parents, we could either compare the two means (as Dr. Woolf recommends) or we could carry out an analysis of variance: with either method the answer would of necessity be identical. Where more than two means have to be compared, an exact algebraic identity no longer obtains; but there can be little question that for general purposes the second procedure is at once the simpler and the more efficient. Thus, the contrast between first degree and second degree statistics turns out to be rhetorical rather than real.

For a convincing answer to the more technical details of Dr. Woolf's criticisms, the reader may be referred to Professor Mather's own paper in the same publication ([33], pp. 103–111). As he remarks, there are certain minor but undeniable defects in the particular researches that Dr. Woolf has discussed; and these reduce the cogency of the conclusions reached. But most of the shortcomings seem to arise, not from the inadequacy of the statistical procedure,

but from inadequacies in the data available or in the planning of the experiments by which the data were obtained.

V. ALTERNATIVE HYPOTHESES

The Inductive Procedure

In conclusion we should like to draw attention to one point that nearly all our critics have overlooked, namely, the logical form of the argument by which we have defended the hypothesis advanced. As in all empirical problems, the whole argument is essentially inductive, not deductive; and it is convergent rather than serial. Dr. Harrison declares that "the chain of Burt's reasoning can be no stronger than its weakest link: so long as any one of his premisses remains open to doubt, the conclusion that hangs on it is liable to fall to the ground, and the reader is left free to choose some other way of explaining the facts". But the conclusion was sustained by several chains, not by one. In the paper on 'Evidence of the Concept of Intelligence', five different proofs were offered—observational, introspective, genetic, neurological, in addition to the mere statistical analysis, which we ourselves regard as no more than confirmatory. Thus the theory we have advanced really rests on what Whewell called 'a consilience of inductions'.

Nor is it sufficient to show that the facts adduced *may* be explained in some other way: the critic must also show that his way is *more probable* than ours. When Dr. Harrison adds that our own explanation is 'never conclusively proved', we readily agree: in the natural sciences, nothing is ever *conclusively* proved. The final inference was put forward not as conclusive or certain, but merely as more probable than any alternative explanation. The only way to rebut this argument is for the critic to produce a constructive hypothesis of his own, and formally demonstrate its higher probability.

What then are the rival explanations that our opponents have to offer? We have searched their pages [16], [21], [26], [33]; and nowhere can we find a clearly defined theory, much less any formal proof, which would show, as Fisher for example showed, that the new theory yields a better fit to the facts observed. However, judging from fragmentary comments that have emerged in the course of the discussion, there would seem to be two more or less plausible alternatives that compete with our own. Most of the critics who, like ourselves, accept the possibility that intelligence is in part inherited would apparently favour some form of the venerable theory of 'blended' inheritance in place of the neo-Mendelian theory of 'multifactor' inheritance. The rest of our opponents seem to hold that inheritance plays only a negligible or at any rate an unimportant part, and ascribe the observable differences between individuals primarily to environmental influence.

1. *Blended Inheritance.* The theory of 'blended inheritance' has the sanction of Darwin's name behind it, and was almost universally accepted until the dramatic rediscovery of Mendel's hypothesis of 'particulate inheritance'. The theory of blending is also the explanation which, by its very simplicity,

naturally commends itself to those who are unfamiliar with modern genetics. The facts most frequently adduced in its support are those of 'regression', which both Galton and Pearson interpreted in terms of blending. Put in explicit shape, the argument runs somewhat as follows. Suppose the general mean of the population to be 100 I.Q. units, and that of the father to be $100 + x$ (where x may be either positive or negative); then, on the hypothesis of random mating, the most probable level of the wife's intelligence will be 100; hence that of the child will be the average of the two, namely,

$$\frac{1}{2}\{(100 + x) + 100\} = 100 + \frac{1}{2}x;$$

thus, as a result of bisexual reproduction, there will be a regression of 0.50. To take account of the effects of assortative mating and remoter ancestry, a slightly more complex version was elaborated by Karl Pearson ([30], p. 456). But the underlying principle remains the same.

Here, however, a mere comparison of means (such as Dr. Woolf recommends) does not enable us to decide between the hypotheses of blending and of multiple factors, since, as we have seen, the value deduced from each of the two theories would (under the conditions stated) be exactly the same. But, when we turn to the second degree statistics, we find a striking difference. For simplicity, let us take the general mean to be zero; and let the deviations of father, mother, and child be x , y , and z respectively, and the correlation between fathers and mothers r . Then (ignoring such complicating influences as dominance and environmental differences) the deviation of the child will, on both hypotheses, be the average of its parents' deviation, viz., $z = \frac{1}{2}(x + y) = \frac{1}{2}(1 + r)x$ ([9], p. 115). If there are N males and N females in each generation, the variance of the adults will be $\sum x^2/N$ and $\sum y^2/N = \sigma^2$ say (since with autosomal factors the difference between the variances of the two sexes would be nil). On the multifactor hypothesis, as we have seen, the variances of the children would remain the same ([9], p. 110); but, on the hypothesis of blended inheritance, the variances of the children would be $\sum z^2/N = \frac{1}{4} \sum (x^2 + 2xy + y^2)/N = \frac{1}{2}\sigma^2$ if mating is random, or $\frac{1}{2}(1 + r)\sigma^2$ if mating is assortative. Thus, the effect of blending is to reduce the inheritable variance by roughly one half in each generation. Now the data both for bodily and for mental characteristics make it clear that no such reduction takes place. On this hypothesis, therefore, we are driven to assume, as Darwin did, that some unknown cause or causes must be continually at work to produce fresh 'spontaneous variations'. But all that we know of the rate and nature of mutation puts any such ad hoc assumption completely out of court.

Accordingly, in the absence of more positive evidence in its support, we may conclude that the hypothesis of blended inheritance has a far lower probability than the hypothesis we have proposed.

2. *Environmental Influence.* Those who hold that the mental differences between individuals are the effect chiefly of environmental conditions commonly follow Watson and the behaviourist school, and draw a sharp distinction between

behaviour and structure. "Differences in structure", we are told, "are, or may be, to a large extent hereditary; differences in behaviour are due almost exclusively to training. . . . Man is, before all others, an educable animal; and conditioning—the simplest form of education—begins so early, even as early as embryonic life, that any differences in hereditary conditions are completely overlaid by the effects of learning". "Give me a dozen healthy infants", says Watson, "and my own specified world to bring them up in, and I will guarantee to train any one to become any type of specialist—doctor, lawyer, artist, beggar-man, or thief" ([39], pp. 104f.). These and similar optimistic claims are, we are told by Dr. Fleming and others, largely borne out by the work of Dr. Stoddard and his colleagues at Iowa ([16], pp. 111f.). As a result, "since the opening of the fifth decade of the present century a greater belief in educability has characterized educational thinking". Nevertheless, Dr. Fleming continues, even so recently as 1931, "it was possible for an official publication to announce" that the psychological evidence "necessitated the provision, not merely of separate classes, but of separate types of schools for pupils of differing capacity"; and she quotes Burt as an "early sponsor" of this view, who has "recently repudiated" it. "It is no longer supposed", she adds, "that children can thus be classified into a small number of discrete types". Children who were formerly classified as 'incurably defective' are now grouped with milder cases of backwardness, and given special training; and even the current doctrine that children who are 'innately bright' can be picked out at 11 plus for transference to grammar schools is "now thought to be an injustice to those who are not".

There is a slight misunderstanding here: Burt has never "repudiated" the evidence for separate schools and classes. But for the rest, we should be in fairly close accord with Dr. Fleming's comments on the administrative changes in current educational practice. Yet that is not because we attach less weight than before to the importance of genetic influences. In any case, this "new emphasis on the powerful effects of environment" is by no means so novel as these writers apparently suppose. It constituted the chief explanation for individual differences offered by the 'philosophic radicals' of the nineteenth century—the founders of the old associationist school which for so long dominated educational theory in this country; and even Darwin at first attributed 'individual as distinct from specific or varietal difference' to 'environmental pressure rather than inborn endowment', until, as he handsomely owned, Galton's book on *Hereditary Genius* 'made a convert' of him¹. The argument which Galton himself used to controvert 'the excessive reliance on environmental explanations' could be turned no less effectively against American and British behaviourists of the present day. Since they admit that man's 'unique educability' is inherited, just like his power of speech, his foetal face, and his lack of a visible tail, is it not extremely probable that innate deviations in the former characteristic (as in all other innate traits) should distinguish different

¹ See the letter quoted in Galton's *Memories of My Life* (1909), p. 290.

members of the race? Why should educability be the sole biological trait which exhibits no individual variation?

How precisely differences in the environment are supposed to act in order to produce the wide differences observable between one individual and another, is hinted at only in the most vague and speculative way. Nor is it altogether clear what exactly its partisans include under the term environment. We are inclined to suggest that there are three main sets of causal process that require to be considered.

(i) *The Acquisition of Knowledge or Skill.* The older British associationist school, and their descendants the American behaviourists, refer almost exclusively to the consequences of what they variously call 'education', 'training', or 'conditioning'. If, they say, the child from a higher socio-economic class generally manifests a higher degree of intelligence, he owes it simply to the fact that he went to a better school or received a better training at home. How then are his superior opportunities supposed to operate? Chiefly, it would seem, though perhaps not solely, by allowing and encouraging him to learn (a) certain items of knowledge and (b) certain forms of skill. These, it is argued, will fully account for his successes or his failures in the tests that the psychometrist employs.

Certainly with individual tests like those of the Terman-Binet scale, the child from the poorer and less cultured home is penalized in questions that depend on the knowledge of recondite words (as in the Vocabulary Test) or on certain semi-scholastic skills (like reading, writing, and drawing). But the child from the 'privileged classes', brought up, as he so often is, in a sheltered home, is just as likely to be inexperienced in such practical matters as giving change, discriminating weights, or finding a lost ball. Here it is to be noted that, as we have pointed out in earlier studies of such tests, the advantage by no means rests exclusively with the child from a superior socio-economic class ([4], pp. 164f).

As for group tests, it is said, they are usually verbal: they require an ability to read words, to understand words, and sometimes to spell them. A child's performance in such tests, so it is maintained, will consequently depend upon his schooling and the cultural background of his home. But this is a non-sequitur. Because the reaction required by the test includes an activity that has been learnt, it does not follow that success at the test is determined, or even affected, by *differences* in what has been learnt. The activity may be so simple and automatic that, at the age when the test is applied, it gives rise to no individual differences whatever. Any psychologist who is adept in the construction of such tests will scrupulously avoid words or forms of scholastic skill that are likely to be harder for some children than for others; and will so arrange the tasks required that the child's success in each turns solely on his ability to solve diagnostic problems calling for no special training and no prior experience. In a few rare cases, where the child has lacked the ordinary opportunities for picking up even the essential rudiments—as with those who have spent their

lives on barges or canal-boats—the test-results may at times be distorted by environmental handicaps. But, when that happens, the blame should rest rather with the investigator who chose an unsuitable procedure than with the test as such. A school psychologist who is attached to the education authority's staff should have ample opportunity and adequate resources for detecting these exceptional cases; and the need for such vigilance, and the systematic use of various means for checking such possibilities, were points that were fully recognized in the days when backward pupils could be nominated for certification as mentally defective. Hence we feel fully convinced that, in our own data, errors of this kind have been almost entirely obviated by the plan we have adopted.

(ii) *Stimulation*. Many of our critics would be willing to accept these various contentions. But they will explain that what they themselves mean by environmental aids are not so much the items of knowledge or skill acquired at school or elsewhere, but rather the general mental stimulation provided by a cultured home. The 'stimulation' is supposed to operate partly by giving the child ample scope and encouragement for exercising his latent powers, and partly by "inculcating a higher motivation and habit of accurate and speedy work". And thus, as Dr. Fleming tells us, "the longer a child remains in unstimulating environments, the lower becomes his performance in tests of what is called intelligence, and the earlier he is transferred to intellectually stimulating situations the higher becomes his measurable response".

This version of the environmental hypothesis owes its popularity to the widespread belief in 'transfer of training'. Intelligence, it will be seen, is conceived as a kind of faculty, which is strengthened by continued practice and liable to atrophy through disuse. Without wholly endorsing this familiar doctrine in its usual guise, we are inclined to agree that certain tests may in some small measure be influenced by the child's acquired habits of alertness, responsiveness, and the like. At the same time, it is our impression that the effects of such 'stimulation' are to be traced quite as often among children from the less cultured classes as among those from better educated families. As one of the headmasters told us: "If anything, your tests seem to favour the 'smart alec' and the future 'spiv' rather than the child from an intellectual home". But in any case the effects are slight as well as disputable; and those who seem to make them part of their hypothesis are avowedly relying on mere surmise, and offer no first hand evidence for their views. In the rare instances in which the examining psychologist suspected any such factors, it would be his duty to allow for differences in attitude or background, and amend his assessment in the light of fuller or better information.

(iii) *Pre- and Post-natal Health*. The environmental conditions to which we ourselves should attach the greatest weight are those affecting the development of the child during the earliest year or two of life, including the period spent *in utero*. In certain cases there can be little question that illness, weakness, or malnutrition during these formative stages has permanently impaired not only

the child's physical health but also the growth of his central nervous system. Moreover, the effects that ensue cannot, by their very nature, be eliminated by changing the type of test or by improving the method of assessment. Even among well-to-do families, the first child, and those born when the mother is advanced in years, seem prone to suffer from pre-natal handicaps and perhaps from the inexperience or the incompetence of the mother in her capacity as trainer and nurse. Some indication of the effects that follow may be gleaned by classifying test-results according to the order of the child's birth: with the majority of tests, as one of us has pointed out in earlier papers, the averages for first born and last born are decidedly lower than those for the intervening children. However, with our own assessments such effects appear extremely small (a matter of less than half an I.Q. point). Moreover, during the last few decades there has been a marked decline in the amount of weakness or ill health attributable to causes of this type: that is strikingly shown by the spectacular drop in the figures for infant mortality. Yet, in spite of these improvements, the test performances of the child population have displayed little or no corresponding change in the course of these years; and certainly there is not a shred of evidence to suggest that the proportion of educationally subnormal pupils—dull, backward, and feeble-minded—has in any way diminished during this period, but if anything rather the reverse.

But the most cogent argument against these versions of the environmental hypothesis is the existence of children, suffering in a severe degree from one or other of the handicaps described—often from all three—who nevertheless have gained high assessments for intelligence, and even high scores on tests of the usual type. Anyone who may happen to have worked among children in the poorer slums of our larger cities forty years ago will recall numerous instances of this kind. In our earlier surveys we found, among school children classified as living in 'very poor' or 'poor' conditions (classes A, B, C, and D as defined in *The Young Delinquent*, p. 68), that as many as 28.4 per cent had I.Q.'s above the general mean (100), 3.1 per cent I.Q.'s over 120, and 0.4 per cent I.Q.'s over 130. All of these came from homes of a deplorably low intellectual status; the majority had received a far more meagre education than the average child—some indeed had made less than half the normal school attendances; and about one-third had suffered from severe malnutrition and ill-health during their earliest years. The obverse of this is equally true. Among children of well-to-do and highly intellectual parents, whose care and treatment have seemed impeccable, much the same proportions have proved to be irremediably dull or hopelessly defective: as evidence, however, such cases are less convincing, since one can never be perfectly certain that some obscure pre-natal disturbance may not have escaped detection.

VI. CONCLUSIONS

Our final conclusion, therefore, may be summed up as follows. First, there can be no question that environmental conditions may influence the

results obtained with intelligence tests of the usual kind. How great that influence is will depend on the style of test employed and the type of child examined: with the ordinary school population, and with tests which have been competently constructed, selected, and administered, we believe that environmental differences are far less influential than genetic, and can as a rule account for no more than 25 per cent of the actual variance. Secondly, we do not deny that environmental factors of the kinds we have enumerated, particularly the last, may to some extent have affected the observable intelligence of individual children to a degree which no refinements in the methods of testing or assessment could possibly eliminate. If so, however, it still seems clear that such effects must be comparatively slight; otherwise it would not be so difficult to secure direct and indisputable evidence to demonstrate their presence. We therefore conclude that environment can have played only a negligible part in determining the actual figures—the variances, the correlation coefficients, the univariate and bivariate distributions—that formed the main basis of our argument.

So far, therefore, this more detailed examination, both of the objections advanced and of the alternative hypotheses put forward, results in no appreciable modification in the views or the methods originally suggested, but, in the main, tends to confirm rather than to confute the conclusions already drawn.

At the same time it is necessary once again to emphasize that these conclusions are strictly limited and avowedly provisional. The most serious defect in our argument has been the lack of any method for assessing the probable margin of error. To meet the doubts expressed by our various critics, it is necessary not merely to demonstrate the *presence* of genetic factors, but also to determine, if possible in quantitative terms, their practical *importance*. Their mere existence Hogben, for example, is ready to acknowledge; but apparently he would prefer to leave it an open question whether their share in the variance amounts to 70 per cent or perhaps to no more than 10 per cent. That, however, would be useless for practical purposes. With our own data, we believe that a rough indication of the margin of error is furnished by the factor we termed 'unreliability'. But a far more rigorous proof is needed. The next step, therefore, should be to plan a full scale inquiry on systematic lines which will permit the application of such tests of significance as are already available (cf. [28]).

And even if our own estimates are accepted as reasonable approximations, it must be frankly recognized that the conclusions so reached are valid only in reference to the particular conditions under which they were obtained. They would not hold good (a) of other mental traits, (b) of different modes of assessment, (c) of a population of different genetic composition, or (d) of a population at a different cultural level: much less would they hold good if there were any subsequent change (e) in the actual distribution of environmental and genetic characteristics, or (f) in the influences affecting their mutual interaction.

But, when all is said, mere statistical inquiries are not enough. Nothing like a decisive pronouncement on the issues we have raised can be looked for until more extensive research has been undertaken in the field of psychogenetics.

In our opinion an adequate understanding of the processes involved can be gained only by carefully planned experiments on lowlier animals, where pure strains can be secured and breeding controlled. When we know more about the genetics of intelligence in animals, then we may be able to construct with greater confidence a more exact hypothesis regarding the transmission of intelligence in man. We therefore fully endorse the contentions of a recent reviewer who declares that "it is a matter of shame and regret that only an amateurish beginning has been made by psychologists in using pure lines in fundamental research in the nature-nurture area" ([36], p. 344). Meanwhile, the practical decisions we all of us have to take—educational, social, political, or matrimonial—are bound to have genetic consequences; and, if the findings here recorded have any validity, those consequences may be so far-reaching that the questions involved should no longer be left to hasty improvisation as each arises, with nothing to guide us but popular beliefs or semi-intuitional guesswork, but should be carefully studied beforehand in the light of an explicit and self-consistent theory—the best we can work out at the moment, even though the evidence supporting it may still be incomplete.

REFERENCES

- [1] BROAD, C. D. (1947). *The Mind and its Place in Nature*. London: Kegan Paul.
- [2] BURT, C. (1912). The inheritance of mental characteristics. *Eugen. Rev.*, IV, 168–200.
- [3] BURT, C., and MOORE, R. C. (1912). The mental differences between the sexes. *J. Exp. Pedag.*, I, 273–284.
- [4] BURT, C. (1921). *Mental and Scholastic Tests*. London: P. S. King.
- [5] BURT, C. (1935). *The Subnormal Mind*. London: Oxford University Press.
- [6] BURT, C. (1937). *The Backward Child*. London: University of London Press.
- [7] BURT, C. (1943). Ability and income. *Brit. J. Educ. Psychol.*, XIII, 83–98.
- [8] BURT, C. (1955). Evidence for the concept of intelligence. *Brit. J. Educ. Psychol.*, XXV, 158–177.
- [9] BURT, C., and HOWARD, M. (1956). The multifactorial theory of inheritance and its application to intelligence. *Brit. J. Statist. Psychol.*, IX, 95–131.
- [10] CAMPBELL, F. (1956). *Eleven Plus and All That*. London: Watts.
- [11] CHESTERTON, G. K. (1922). *Eugenics and Other Evils*. London: Cassell.
- [12] DARLINGTON, C. D., and MATHER, K. (1949). *The Elements of Genetics*. London: Allen and Unwin.
- [13] DAVID, P. R., and SNYDER, L. H. *Genetic Variability and Human Behaviour*. Ch. III in *Social Psychology at the Crossroads*. (Ed. M. Sherif.) Harper.
- [14] DUNN, L. C., ed. (1941). *Genetics in the Twentieth Century*. New York: Macmillan.
- [15] FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edin.*, LII, 399–433.
- [16] FLEMING, C. M. (1948). *Adolescence: Its Social Psychology*. London: Kegan Paul.
- [17] GLASS, D. V. (1954). *Social Mobility in Britain*. London: Longmans and Green.
- [18] HALDANE, J. B. S. (1938). *Heredity and Politics*. London: Allen and Unwin.
- [19] HALDANE, J. B. S. (1946). The interaction of nature and nurture. *Ann. Eugen.*, XVI, 37–46.
- [20] HAZEL, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, XXVIII, 476–490.
- [21] HEIM, A. W. (1954). *The Appraisal of Intelligence*. London: Methuen.

- [22] HOGBEN, L. (1931). *Genetic Principles in Medicine and Social Science*. London : Williams and Norgate.
- [23] HOGBEN, L. (1933). *Nature and Nurture*. London : Williams and Norgate.
- [24] LEWIS, D. G. (1956). The normal distribution of intelligence : a critique. *Brit. J. Psychol.* (In press.)
- [25] LUSH, J. L. (1947). Family merit and individual merit as bases for selection. *Amer. Nat.*, LXXXI, 241-261, 362-379.
- [26] MADDOX, H. (1956). Nature-nurture balance sheets. *Brit. J. Educ. Psychol.* (In press.)
- [27] MARTIN, F. M. (1947). Home background and selection for secondary education. *Eugen. Rev.*, XLVIII, 195-202.
- [28] MATHER, K. (1949). *Biometrical Genetics: The Study of Continuous Variations*. London : Methuen.
- [29] MILES, T. R. (1957). On defining intelligence. *Brit. J. Psychol.* (In press.)
- [30] PEARSON, K. (1900). *Grammar of Science*. London : A. and C. Black.
- [31] PEARSON, K. (1906). *Nature and Nurture*. London : Dulau.
- [32] PEARSON, K., and LEE, ALICE (1903). On the laws of inheritance in man. I. The inheritance of physical characters. *Biometrika*, II, 357-462.
- [33] REEVE, E. C. R., and WADDINGTON, C. H. eds. (1952). *Quantitative Inheritance*. London : H.M. Stationery Office.
- [34] SIMON, B. (1953). *Intelligence Testing and the Comprehensive School*. London : Lawrence and Wishart.
- [35] SMITH, FAIRFIELD (1937). A discriminant function for plant selection. *Ann. Eugen.*, VII, 240-250.
- [36] STONE, C. P. (1947). Methodological resources for the experimental study of innate behaviour as related to environmental factors. *Psychol. Rev.*, LIV, 342-347.
- [37] THOMSON, G. H. (1948). *The Factorial Analysis of Human Ability*. London : London University Press.
- [38] VERNON, P. E. *et al.* (1950). *ap. The Advancement of Science (Brit. Assoc. Ann. Rep.)*. London : Murray.
- [39] WATSON, J. B. (1931). *Behaviorism*. London : Kegan Paul.
- [40] WHETHAM, D., and C. D. (1909). *The Family and the Nation*. London : Murray.
- [41] ROYAL COMMISSION ON POPULATION (1941). *Memoranda presented to the Commission*. London : H.M. Stationery Office.

BOOK REVIEWS

The Organization of the Cerebral Cortex. By D. A. SHOLL. Methuen, 1956. Pp. 16+125. 18s.

This is a book which should be read not only by the neurologist but by every serious student of psychology. Dr. Sholl's main contention is expressed in his concluding words: "whether the cortex is studied by the anatomist, the physiologist, or the psychologist, the model employed should be based on the concept of probability". It must therefore be derived from statistical hypotheses and be discussed in statistical language.

He begins with a review of the chief histological and experimental methods available, and illustrates the various microscopic techniques with a series of photographs revealing a high degree of technical skill. On the basis of his own investigations he then criticizes the familiar maps of the cortex which are reproduced in nearly every textbook and are cited as demonstrating structural differences between various regions of the brain corresponding to their supposed differences in function. As Dr. Sholl points out, even the best of such maps were the outcome of subjective criteria and manifestly influenced by the preconceived ideas of the cartographers, particularly their "rather crude psychological theories". Few recognized "the enormous variability that exists between individual human brains". Many of the differences, as Bok showed twenty years ago, arise merely from irregularities of growth, differences in local blood supply, and above all the varying amount of deformation produced by the curving surface of the brain.

Dr. Sholl himself holds that the characteristic chiefly calling for study is not the type of nerve cell or the predominance of fibres, but rather "the pattern of connectivity". This is largely indicated by the manner and the extent of the branching which the axons and dendrites exhibit. These he relates to other aspects of cerebral organization; and is thus led to a most suggestive discussion of the 'quantification of neuronal connectivity'.

To the psychologist the most interesting chapter in the book is that which describes "the mode of operation of the cerebral cortex". The *ad hoc* explanations put forward by Pavlov, and the somewhat naive theories deduced from the behaviour of ingenious machines like Ashby's homeostat and Grey Walter's mechanical tortoises, are criticized in some detail. Of the two alternative analogies which have dominated most of the older theories—the telephone switchboard with its precise connections, and the field theory treating the cortex as a statistically homogeneous medium, Dr. Sholl inclines, if anything, towards the second: "neural integration must consist of complex patterns of interaction, not of conduction over relatively isolated paths".

In his discussion of mental abilities like 'memory' or 'intelligence', he unhesitatingly rejects the earlier types of hypothesis that tried to allocate them to 'circumscribed regions'. 'Intelligence', he holds, like 'retentivity', must be a general attribute of cortical activity, resulting from its over-all mode of organization.

The evidence for these various conclusions is lucidly expounded, and supported by quantitative determinations. However, as he explains, "the statistical mode of description has been adopted, not because of the technical difficulties involved in the study of individual neurons, but because it is the appropriate method to adopt for the study of a system whose interactions can only be described in terms of probability".

CYRIL BURT.

Psychological Tests and Personnel Decisions. By L. J. CRONBACH and G. C. GLESER. University of Illinois Press, 1957. Pp. xii+165. \$3.00.

Most of those who have written on the theory of mental tests have assumed that the efficiency of any selection procedure may be judged exclusively in terms of the validity coefficients, regardless of the decisions to be taken. Education committees, however, especially in the early days of mental testing, were apt to inquire whether the *increased* information so gained would justify the additional *cost*. Dr. Cronbach has set out to answer the questions thus raised by developing what Wald has taught us to call 'statistical decision theory'. The book which he and his collaborator have produced contains an admirable account of the construction and use of mental tests from this highly practical standpoint. It should be in the hands of all who work in the fields of educational psychology or personnel selection.

W. R. NORTON.

OCCUPATIONAL PSYCHOLOGY

Editor : ALEC RODGER

April, 1957

Volume 31, No. 2

Knowledge of Results and 'Skilled Performance'

JOHN ANNETT and HARRY KAY

Interrelation of Attitudes and Communications in a Sub-Divided Working Group

R. A. M. GREGSON

The Effect of Initial Pointer Position on Display-Control Relationships

D. H. HOLDING

Simplifying the Controller's Task through Display Quickening

FRANKLIN V. TAYLOR

The Transition from Work to Retirement (1)

MARGARET PEARSON

Training in Industrial Skills: Opinions of Trainees

D. C. S. WILLIAMS

Training in Perceptual Skills

MARY D. ALLAN

Book Reviews

Other Books Received

Annual Subscription 30 shillings

National Institute of Industrial Psychology, 14 Welbeck Street, London, W.1

HUMAN RELATIONS

Contents of Vol. X, No. 2

ARTHUR N. TURNER. Foreman, Job, and Company

JAMES SPILLIUS. Natural Disaster and Political Crisis in a Polynesian Society: an Exploration of Operational Research. II

ARNOLD S. TANNENBAUM, ROBERT L. KAHN. Organizational Control Structure: a General Descriptive Technique as applied to Four Local Unions

RONALD TAFT. A Psychological Model for the Study of Social Assimilation

ALAN RICHARDSON. The Assimilation of British Immigrants in Australia

ERIKA CHANCE. Mutual Expectations of Patients and Therapists in Individual Treatment

PAMELA BRADNEY. The Joking Relationship in Industry

30s. per annum or 8s. 6d. per issue

TAVISTOCK PUBLICATIONS LIMITED

2 Beaumont Street, London W.1

CONTENTS

DUNSDON, M. I., AND ROBERTS, J. A. F.	A Study of the Performance of 2,000 Children on Four Vocabulary Tests	<i>Page</i> 1
BERNYER, G.	Psychological Factors: their Number, Nature, and Identification	17
STUART, A.	The Comparison of Frequencies in Matched Samples .	29
BURT, C., AND HOWARD, M.	Heredity and Intelligence: A Reply to Criticisms .	33
BOOK REVIEWS		64

JOURNAL COMMITTEE

CYRIL BURT *Editor*

CHARLOTTE BANKS and ALAN STUART *Assistant Editors*

J. W. WHITFIELD *Managing Sub-editor*

and

B. M. FOSS
M. HAMILTON
A. HERON

J. S. SMALL
J. SUTHERLAND
F. W. WARBURTON

THE
BRITISH JOURNAL
OF
STATISTICAL
PSYCHOLOGY

EDITED BY
CYRIL BURT



WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ALAN STUART
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN	E. A. PEEL
M. S. BARTLETT	L. S. PENROSE
W. G. EMMETT	J. FRASER ROBERTS
M. G. KENDALL	A. RODGER
D. N. LAWLEY	W. STEPHENSON
E. S. PEARSON	P. E. VERNON

Managing Sub-editor J. W. WHITFIELD

Printed and Published by
TAYLOR & FRANCIS LTD.

RED LION COURT, FLEET STREET, LONDON, E.C.4

Price 20s. per part (U.S.A. \$3.50)

Subscription 30s. 6d. per volume (U.S.A. \$4.50)

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Statistical Psychology is issued by the British Psychological Society. The subscription price per volume is 30s. net for Members of the Society and 30s. 6d. (post free) for non-members. The subscription price in the U.S.A. is \$4.50 (post free). Members of the Society should send their subscriptions to THE SECRETARY, British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1: non-members to Messrs. TAYLOR & FRANCIS, LTD., 18 Red Lion Court, Fleet Street, London, E.C.4. Until further notice the Journal will be issued in two parts each year, in May and November.

Papers for publication should be sent to Sir CYRIL BURT, 9 Elsworthy Road, London, N.W.3. Authors will receive 25 copies of their articles free; extra copies may be purchased provided the order is given when the proofs are returned. Contributors should indicate whether they wish their reprints to be bound in covers. The printers will state current prices when sending out proofs.

Books for review should be sent to the Editor, advertisements to the Secretary, British Psychological Society.

The Society also issues quarterly *The British Journal of Psychology* and *The British Journal of Medical Psychology*: subscriptions, 60s. per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society publishes, jointly with the Association of Teachers in Colleges and Departments of Education, *The British Journal of Educational Psychology*: for this the subscription, £1 per annum, should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2. Members of the Society receive the foregoing journals on special terms; enquiries should be addressed to the Secretary, British Psychological Society.

Papers for publication in *The British Journal of Psychology* should be sent to Professor James Drever, Department of Psychology, the University, Edinburgh; those for publication in *The British Journal of Medical Psychology* to Dr. Joseph Sandler, 96 Portland Place, London, W.1; those for publication in *The British Journal of Educational Psychology* to Professor P. E. Vernon, London University Institute of Education, Malet Street, London, W.C.1.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in previous publications of this Journal. Each article should be headed either with a series of numbered headings in italic, corresponding with the cross-headings of the several sections, or with a brief abstract: (the latter procedure is preferred). Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. (particularly in the case of tables and references). In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. II, pp. 21-73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, XL, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Manuscripts involving many tables and numerous equations can as a rule only be accepted if the authors are willing to pay for the additional cost.

In future authors will receive two complete slip-proofs of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

PROFESSOR LEWIS M. TERMAN

32

Readers will have heard with deep regret of the death of Professor Terman. In this country he was perhaps best known for his revisions of the Binet-Simon scale. In the days when mental testing was still looked at askance by the majority of psychologists and educationists, he perhaps more than anyone succeeded in giving it a scientific status and putting it into practical shape. As he himself points out, he had no wish to contribute to statistical theory or to the principles of mental measurement, but only to make use of them. And by his own example he became one of the most effective advocates of metrical techniques, and did much to convince both teachers and administrators of the practical value of the new statistical procedures.

His origins, like that of so many eminent scientists, were curiously mixed. He was, as he himself puts it, "Scotch, Irish, Welsh, German, and French". He was born in a country district about seventeen miles from Indianapolis on January 15, 1877. From data given by J. M. Cattell in *American Men of Science*, we learn that in a group of 1,000 scientists Terman was the only one belonging to a family of twelve or more. He was in fact the twelfth of fourteen children; and his father was one of ten. The Termans, Tearmans, or Turmans were a roving and prolific family; and his branch of it had apparently migrated Westward from Virginia. Most of them, like his grandfather and father, were hardworking and prosperous farmers. Nevertheless, the household library contained something like two hundred volumes: and the young Lewis apparently found books more interesting than the farmyard. Among his favourites he mentions Dickens, the *Encyclopaedia Britannica*, and *Peck's Bad Boy*.

There was, he says, nothing in his early environment that could have "conditioned him specifically in favour of psychology". However, "almost as far back as I remember, I seem to have had rather more interest than the average child in the personalities of others, and to have been impressed by those who differed from the common run". His elementary education was acquired in "a little red one-room school-house, which could not boast a single library book". From eleven to eighteen he worked on the farm for about six months each year; and "any mental development that occurred during this period must have been due to maturation rather than to intellectual stimulation".

At school, and later on at college, he preferred linguistic subjects to mathematics and physical science. In this respect he notes a marked contrast between himself and another celebrated member of his class, Arthur Banta, the biologist. "Banta," he says, "was less bookish and less interested in the traits of people, but far more interested than I in plants, animals, and rocks." As it happened, both his school and his neighbourhood contained few boys of his own age. Most of them were older. And this, he thinks, may have provided ground for "the inferiority attitude and tendencies to introversion," which even in adult years he still revealed in the Bernreuter tests. A disposition of this type he

S. P. Bureau of Ednl. & Psyl. Research
(S. C. E. R. I.)

Bureau of Ednl. & Psyl. Research
(S. C. E. R. I.)

thinks may be somewhat characteristic of the psychologist, since it inclines a man towards introspection, and makes him secretly curious about the abilities and motives of those around him. Even before he reached his teens, he became fascinated by after-images, by the flight of colours, and by the connections between one idea and the next. He mentions an early discovery, which recalls a similar experience of Tennyson. "When about eleven or twelve years of age, I hit on the fact that, by repeating a formula, I could lose my sense of identity and my orientation in time and space: my method was to gaze fixedly on something, and repeat: 'Is this me?' until all distinction between subjective and objective was lost in a kind of mystical haze."

While he was still only ten years old, "a pedlar who sold books on phrenology stopped at the farm for a night, discoursed on his science, and felt the bumps of each of us . . . When my turn came to be examined," says Terman, "he predicted great things of me". That, he believes, may have stimulated him to greater efforts. And the interest in phrenology which was thus implanted served as his first introduction to the science of individual differences.

When he reached the age of fifteen, he went to a 'normal college' to prepare for teaching. Of the various psychological textbooks that he studied, he "liked Sully" best. James' *Principles* he did not discover until later. But, he adds, "of all the founders of modern psychology my greatest admiration was for Galton". At twenty-one he became principal of a township high school, where he "taught the entire curriculum to about forty pupils". A year later he married. His wife relates how he once told her that his "interest in their baby had determined him to become a psychologist". In 1901 he went to Indiana University; and a couple of years afterwards to Clark University—in those days the Mecca for almost every aspiring young psychologist in America. He was one of a group of brilliant students, twelve of whom later became sufficiently distinguished to be included in the American *Who's Who*.

For a while he continued under "the hypnotic sway of Stanley Hall". But he soon began to feel that the value of Hall's questionnaire procedures was getting exhausted. His own paramount interest still remained centred on individual differences, and more especially on methods of studying the more extreme cases—the gifted and the mentally deficient. He recalls two essays which he prepared for seminars—on 'The Great Man Theory' and on 'Degeneracy'. Instead of the stock questionnaire, he began to consider the possibilities of mental tests. He thinks that, as with Thorndike, his mind may have turned in this direction partly because he "had no aptitude for, and little interest in, the laboratory techniques of the experimentalist of the Wundt-and-Titchener type". Stanley Hall strongly disapproved. Terman, however, started to read whatever he could find on the subject, whether in English, French, or German: he mentions more particularly Galton, Binet, Stern, Cattell, and above all Thorndike. One thing, he found, was badly needed and that Clark University did not offer, namely, "instruction in statistical methods": "a year with Thorndike would have been an untold boon".

Unfortunately his studies were interrupted by a pulmonary haemorrhage and the discovery of a mildly active tuberculosis. On recovering, he was offered the chair of Child Study and Pedagogy at Los Angeles. Here he still had to nurse his health. But four years later in 1910 he was appointed professor of educational psychology at Stanford.

The first fruit of his lectures at Stanford was the small textbook on *The Hygiene of the School Child*. Huey, however, who with Gesell had for long been one of his closest friends, urged him to start an investigation on the value of the Binet scale of tests which had already been used for work on defectives by Goddard, Kuhlmann, and others. Accordingly, in collaboration with one of his graduate students, H. G. Childs, Terman began an experimental study of the 1911 scale which eventually led to the publication of *The Measurement of Intelligence* in 1916.

It was during the same period that Burt started his investigations on much the same problem; and it is curious to observe how both of them, working to begin with, it would seem, with no knowledge of each other's interests, studied the same writers and adopted much the same lines of approach. Both Terman and Burt record their indebtedness to Galton's work in Britain, Meumann's in Germany, and Binet's in France. Owing, however, largely to the influence of Pearson, Burt became interested in the quantitative problems; Terman, on the other hand, related that his "own interest in mental tests at that time was more in their qualitative than in their quantitative aspects". But both appear to have reacted in much the same way to Spearman's methods and conclusions. Terman describes how, on reading Spearman's paper in the *American Journal of Psychology*, he was greatly impressed with "the dogmatic tone of the author, the finality with which he disposed of everyone else, and his one-hundred-per-cent faith in the verdict of his mathematical formulae"; but, he adds, "the conclusion to which it all led, namely, that there is 'a correspondence between what may provisionally be called General Discrimination and General Intelligence which works out with great approximation to one or absoluteness', seemed to me as absurd then as it does now".

The book on *The Intelligence of School Children* and the monograph on the Stanford Revision data were "by-products of the work that led to *The Measurement of Intelligence*". As in Britain so in America, "many of the old-line psychologists regarded the whole test movement with scorn"; "hence", says Terman, he himself "was a little surprised that his publications were so favourably received". Indeed, even at the time, he "imagined the revision would probably be displaced by something better". And so eventually it was, namely, by 'the new revised Stanford-Binet tests' produced by Terman in collaboration with Miss Merrill in 1937. British investigators who either then or later on were trying to modify or standardize the original Binet-Simon scheme, so as to render it more suitable for use in this country, will always remember with gratitude the generosity with which Terman permitted them to borrow or adapt either of his own revisions.

In the meantime, as soon as the first world war was over, Terman devoted his attention to the standardization of group tests of ability and scholastic tests for assessing educational achievements. At length a generous grant from the Commonwealth Fund enabled him to undertake the researches that he had always had in mind on 'the psychology of gifted individuals'. The result was the three large volumes, published between 1924 and 1930, under the title of *Genetic Studies of Genius*. A few years later, he turned to what at first sight appears a somewhat unexpected field of investigation, although, as he himself records, the problem had been germinating in his mind for many years as a result of his comparative studies of boys and girls during their years of growth and schooling. The outcome was the work on *Sex and Personality*, based largely on his well known 'Masculinity-and-Femininity Index', and published in collaboration with C. C. Miles.

In his last publication, his Walter Bingham Lecture (1954), he reverted once again to the topic nearest to his heart, 'The Discovery and Encouragement of Exceptional Talent'. The problem of the gifted individual, he maintains, still fails to receive the attention it merits; and for this reason he is "inclined to be pessimistic about present trends in democracy". His closing words are well worth quoting. "What I especially want to emphasize," he says, "is that the evidence on early mental development of historical geniuses and the follow-up of gifted subjects selected by mental tests point to the conclusion that capacity to achieve beyond the average can be detected early in life by a well-constructed ability test heavily weighted with the *g* factor."

It would be presumptuous to attempt any review of Terman's great and lasting contributions to psychology. But he himself has summarized the salient points. "If I am remembered after my death," he writes, "it will probably be in connection with my studies of gifted children, the construction of mental tests, and the psychology of sex differences. I think that I early saw, more clearly than others, the possibilities of mentality testing, and have succeeded in devising tests that work better than their competitors, and, by the application of test methods, have added to the world's knowledge of exceptional children." He believes that "the revival of associationism by 'bond' psychology has run its course: the Watsonian brand of behaviourism is a cult; and a denial of heredity on the basis of a few minor experiments on the conditioning of infants ridiculous; the major differences between children of high and low I.Q., and the major differences in the intelligence test scores of certain races, will never be fully accounted for on the environmental hypothesis." These conclusions plainly have an importance beyond the bounds of psychology; and, if they are gradually winning acceptance, that is largely due to the long and tireless labours of Professor Terman himself.

WILLIAM B. LEWIS.

THE RELATIONS OF THE NEWER MULTIVARIATE STATISTICAL METHODS TO FACTOR ANALYSIS

By HAROLD HOTELLING
University of North Carolina

Abstract. A survey of developments in multivariate analysis during the last thirty years shows that some, though not all, of the purposes for which factor analysis has been used may now be better accomplished by other procedures, e.g. by regression, multiple correlation, the study of relations between two sets of variates, or of the dimensionality of a set of variates.

To determine whether two or more groups of persons differ significantly in their mean values or their covariance matrices, the most appropriate procedures consist of methods of multivariate analysis of variance. Such methods have increased rather than diminished the advantages of using an external criterion instead of making a purely internal analysis.

The problem of estimating the dimensionality of a continuous multivariate population by means of a sample admits a simple and exact answer: the rank of the population is equal to that of the sample, provided the number of degrees of freedom among the individuals in the sample exceeds the number of the variates. If, however, the problem is re-interpreted so as to imply that each observation is the sum of a real part and a random error, then the hypothesis that the real part has a rank smaller than the number of variates can be tested only when a proper estimate of the errors is available, i.e. by making suitable replications. The appropriate method is again multivariate analysis of variance rather than factor analysis. In a third form of the problem the object is, not to ascertain a true dimensionality, but to assess the error resulting from discarding (on practical grounds) a dimension that may really exist. Here the problem is somewhat indeterminate: a practical expedient may involve a special form of factor analysis, viz. the calculation of the least, not the greatest, latent roots.

Thus, in many cases in which they have hitherto been used, factor analyses of the usual kinds are inferior to other procedures: nevertheless, the results of such analyses may have heuristic and suggestive value, and may uncover hypotheses which are capable of more objective testing by other methods.

I. RECENT DEVELOPMENTS IN MULTIVARIATE ANALYSIS

Factor analysis in psychology, and the more general study of multivariate analysis in statistics, have developed side by side and to a great extent independently. Owing to lack of adequate communication between the two groups of research workers—each spread geographically, but having few contacts with the other—psychologists have often employed statistical methods demonstrably inferior to those worked out by mathematical statisticians, while the

latter have often been unaware of the highly interesting activities and stimulating problems of psychologists. The following paper outlines achievements in multivariate statistical analysis during the last quarter-century, suggests uses in psychology of the newly discovered methods, and tries to mark out a boundary between the cases in which these should supersede methods involving factor analysis and those in which factor analysis is still the most useful tool. It will be seen that some, but not all, of the things people have undertaken to do by factor analysis can be done better in other ways. The methods here considered have all been derived primarily for use with multivariate normal distributions; in other cases they still have certain descriptive, suggestive, and approximative values.

(i) A primary tenet of modern statistics is that inferences must be supported by known probabilities to keep under control the effects of errors in measurement and in sampling. (ii) In selecting statistical methods a second guide-post is that they must maximize the probabilities of right conclusions and minimize the probabilities of wrong. This position, which was illustrated by Gauss' espousal of the mean square error as against less efficient measures of dispersion, has been strongly emphasized in a series of influential papers by R. A. Fisher, beginning with his memoir on 'The mathematical foundations of theoretical statistics' [5]. It implies that the *choice* of methods of descriptive statistics is logically posterior to the *study* of statistical inference, though the *practice* of descriptive statistics comes normally before the inference. (iii) Thirdly, the statistician has abandoned his old passive role of merely analysing whatever was put on his doorstep by other investigators or by chance, and now has much to say about how experiments should be designed to make inferences from them valid and the whole process economical.

The first principle, calling for known probabilities, has stimulated much ingenious research in methods of finding probabilities; but it has also done something more fundamental in stipulating that the probabilities are to be known numbers wherever possible, and not merely functions of unknown parameters that can only be estimated. It was to satisfy this stipulation that the 'Student' distribution was introduced, and later on, the analysis of variance and the exact distributions of the ordinary correlation coefficient, the multiple and partial correlation coefficients, the multivariate generalization of the Student distribution, the multivariate forms of analysis of variance, and canonical correlations. Here again it was R. A. Fisher who obtained the first of these exact distributions, and led the way in showing how to put them to practical use.

The multivariate normal distribution in p variates, even with zero means, has $p(p+1)/2$ independent parameters; and the distribution of an arbitrarily chosen 'statistic', or function of the observations, may involve all of them. If p is at all large, the number of these parameters is very large indeed. When, as is usual, none of them is known apart from the observations, the control of error probabilities becomes hopelessly difficult unless the statistics to be used are carefully chosen so as to keep down the number of these unknown quantities,

which I have termed 'nuisance parameters', entering into the most relevant probabilities.

This need of keeping down the nuisance parameters limits drastically the statistical methods appropriate for use with multivariate normal distributions, going beyond the second, or 'efficiency' tenet, which leads to the use of sufficient statistics where they exist, as they do the multivariate normal distribution. The sufficiency criterion tells us here that we should use no statistic that is not a function of the sample means, variances, and covariances. The new criterion tells us further that among these functions we should seek to use only those in whose distributions the nuisance parameters can be kept under control and ideally are non-existent. Many of the powerful methods discovered in the last three decades are based on this principle.

The first example is the multiple correlation coefficient, whose general distribution in samples from a multivariate normal population was discovered by Fisher [6]. It looked at first as if the distribution would involve all the covariances as independent parameters, and therefore be excessively complicated and difficult to determine as a function; moreover, even if the distribution function of R were determined, it would be useless if it involved all these many nuisance parameters, since they are themselves unknown. Fortunately the situation was saved by the invariance of R under all non-singular linear transformations of the predictors, or 'independent variables'. Such a transformation can always be found that will make the predictors mutually uncorrelated, and will also make all but one of them uncorrelated with the predictand. Thus in the transformed system there is only one correlation that can be different from zero, and this equals the population multiple correlation coefficient ρ . Hence the distribution of R involves only ρ ; and Fisher, using this invariance property and some ingenious geometry, with other tricks, found the exact probability density.

The great principle illustrated by Fisher's success with the multiple correlation coefficient is the possibility of eliminating nuisance parameters and simplifying exact probabilities when the statistics used are invariant under all non-singular linear transformations of the variables, or even under an extensive class of such transformations such as the orthogonal. This puts a premium on the use of invariant statistics, and points to the selection of such statistics to meet various needs. The criterion of invariance has been influential ever since Fisher's 1928 paper in the development of new methods of multivariate analysis.

II. THE GENERALIZATION OF STUDENT'S DISTRIBUTION

The manifold uses of the Student distribution in one dimension, for example in testing the significance of means and differences of means, and in laying down confidence intervals with known probabilities, had long required generalization to two or more dimensions, and various unsuccessful attempts had been made in this direction. Notable was the 'Coefficient of Racial Likeness', proposed by Karl-Pearson [13], which, like other such efforts, foundered on the nuisance

parameter rocks. It was the sum of squares of differences between corresponding means in two samples, with each square divided by a variance estimate drawn from the data of both samples. Each term represented some measured trait. The trouble was that this statistic, which was used with enthusiasm in the 1920's, had a distribution that was not only impossibly hard to find, but involved all the unknown correlations and ratios of true variances among the variates.

The remedy eventually found for this trouble was the replacement of the simple diagonal matrix of the quadratic form in the differences of the means by a matrix inverse to that of the sample covariances among these differences. The square root T of this quadratic form is a direct generalization of the t of the Student distribution, and for p variates has all the uses that t has for one. For example, it may be desired on the basis of a random sample of N individuals from a p -variate normal distribution to define a region in a space of p dimensions which will have a probability 0.95 of covering the point whose coordinates ξ_i are the true but unknown mean values of the variates in the population. This may be accomplished as follows. Defining $n = N - 1$, $s_{ij} = S(x_i - \bar{x}_i)(x_j - \bar{x}_j)/n$, where the summation is over the sample, we invert the p -rowed square matrix $[s_{ij}]$ to get $[l_{ij}]$, and consider the quadratic form

$$T^2 = N \sum \sum l_{ij}(\bar{x}_i - \xi_i)(\bar{x}_j - \xi_j).$$

The value of this expression is invariant under all non-singular linear transformations among $x_1, \dots, x_p$; and this fact makes it possible to obtain its distribution exactly and in a relatively simple form, easily transformable to the incomplete beta function or variance ratio distribution. Since tables of the latter are now widely available, it is convenient to obtain from them the numerical value F having probability 0.05 of being exceeded, with $n_1 = p$ and $n_2 = n - p + 1$ degrees of freedom, and then to compute

$$T^2 = npF/(n - p + 1).$$

Equating this value of T^2 to the quadratic expression above gives the equation of the boundary of the required confidence region in the space of $\xi_1, \dots, \xi_p$. I have given the full derivation, valid both for this and for various other uses, in the *Annals of Mathematical Statistics* [8], and have discussed and illustrated these methods further in *Statistics and Mathematics in Biology* [11]. The general formula for T^2 is

$$T^2 = \sum \sum l_{ij} \delta_i \delta_j,$$

where the δ 's have the same covariance matrix as the x 's and zero means, and are independent of the sample covariances. Thus in testing the significance of the differences between the means in two independent p -variate random samples from normal distributions with the same covariance matrix, we put

$$\delta_i = \frac{\bar{x}_i - \bar{x}'_i}{\sqrt{\left(\frac{1}{N} + \frac{1}{N'}\right)}},$$

where the sample numbers are N and N' . The distribution is the same as before, with $n = N + N' - 2$; this n is also the denominator of each of the covariance estimates, which are formed by pooling corresponding product sums in the two samples.

S. S. Wilks [17] studied certain generalizations of the analysis of variance to p dimensions, using determinants and ratios of determinants having invariance properties under linear transformations. Other types of multivariate analysis of variance have since been investigated by S. N. Roy, P. L. Hsu, M. S. Bartlett and others. These are suitable, among other things, for testing significance of differences among three or more groups, representing for example different treatments, and generalize the use of T as the ordinary univariate analysis of variance generalizes the use of the Student t .

If in the formula above for T^2 in terms of the δ_i the latter do not all have zero expectations, the distribution is of a more complicated form which was obtained by R. C. Bose and S. N. Roy [4]. As in the case of the multiple correlation coefficient, the invariance property made it possible to show that only one parameter enters into the distribution, and enabled Bose and Roy to obtain the probability element as a multiple of a confluent hypergeometric function. When the δ_i are suitable multiples of the differences between corresponding sample means in two populations, the parameter Δ^2 in the distribution of T is a quadratic form of the same type as T^2 , with population instead of sample means and covariances; Δ may then be called the figurative distance between the two populations if these have the same covariance matrix, and possesses the properties of distance, including the triangular inequality. The Pearson Coefficient of Racial Likeness should be replaced by T^2 for samples, by Δ^2 for populations. Much empirical work using this modernized coefficient of racial likeness has been carried out in Bengal under the leadership of P. C. Mahalanobis. For unequal covariance matrices in the two populations and matched samples, T may be used in an obvious way generalizing Student's celebrated method of comparing hours of sleep gained by using two isomers of hyoscyamine hydrobromide.

III. THE RELATIONS BETWEEN TWO SETS OF VARIATES

A different kind of invariance is involved in the determination of relations between two sets of variates, as in my paper in *Biometrika* [9]. One set might consist for example of scores in several mental tests given to army recruits, and the other of scores on various kinds of subsequent performance. The determination of a linear function of the test scores and a linear function of the later performance scores, such that the correlation between these two linear functions is as great as possible, leads to a determinantal equation whose greatest root is the required maximum correlation. The other roots determine other matching pairs of linear functions, such that each of these linear functions of the variates in one set is uncorrelated with all but its own partner in the other set. These functions are the canonical variates; and the correlations between matching pairs

are the canonical correlations between the two sets of variates. If the canonical correlations are all zero, the two sets are completely independent; and it is useless to predict performance by means of any of the tests used. If one of the canonical correlations is unity, there is something about the performance, represented by a certain linear function of the performance measures used, that can be predicted perfectly by means of the tests. If two canonical correlations are unity, there is a whole one-parameter family of types of performance that can be predicted perfectly by means of these tests.

A use sometimes made for factor analysis in the past is in testing for the relations between two sets of variates. One study of the relations of character with mentality, for example, used seven mental tests and seven estimates of character traits by acquaintances. A standard factor analysis technique was applied to the mental measurements, and independently to the character scores, so as to get seven mental factors and seven character factors. A name was applied to each of the fourteen factors thus found; and a plausible matching of character factor with mental factor was arranged so as to get seven pairs. The correlations within the pairs thus calculated were then computed, and judged in each case insignificant.

This kind of use of factor analysis should clearly be superseded by an examination of canonical correlations between the two sets of variates. The largest canonical correlation will be larger than all the correlations between the two sets actually found, and has a better chance of being found significant and thus demonstrating the existence of a relation between the two sets if one actually exists. The other canonical correlations, in decreasing order of size, and the corresponding canonical variates, will help to elucidate the nature of any relations that may exist between character and mentality. The separate factor analyses of the two sets were really quite useless for this purpose.

A simpler illustration of the need at times to resist the temptation to perform factor analyses is provided by ordinary regression problems, in which a least-squares formula for a single variate y is to be determined by variates $x_1, \dots, x_p$. This is indeed a special case of two sets of related variates, in which the number in one set reduces to unity, the maximum canonical correlation to the ordinary multiple correlation, and the other canonical correlations to zero. If as a preliminary one performs a complete factor analysis on $x_1, \dots, x_p$ —let us say a resolution into principal components—the resulting ‘factors’, or components of the complex will have exactly the same multiple correlation with y as did the original x ’s; and a regression equation for y with them as predictors will for every individual give the same prediction as if the factor analysis had not been performed. The labour of solving the normal equations will be greatly reduced if orthogonal factors are used, but this saving is trifling as compared with the heavy labour of the factor analysis itself if p is large. (Here of course exceptions must be made, including the possibility of electronic computers that will reduce all this labour virtually to zero; also the remote chance of all the correlations among the original variates being zero.) But if not all the principal components are used

in the regression analysis, the results will be virtually affected. If even the last principal component, the one making the least contribution to the total variance, is omitted, both the regression function and the multiple correlation coefficient may be drastically different.

To see this, consider the case of two predictors, x_1 and x_2 , which with the predictand y are represented by three vectors through the origin in the sample space. The multiple correlation is the cosine of the angle θ made by the y -vector with the plane of the two x -vectors, and the regression values are represented by the vector $\mathbf{Y}$ within this plane making a minimum angle with y , and therefore the orthogonal projection of y on the plane. Replacement of the two predictors by their two principal components or other 'factors' is equivalent to rotating each of the x -vectors within their common plane, a process that does not affect the regression vector $\mathbf{Y}$ nor the angle θ , since these depend only on the plane and on the external predictand (or criterion) y . But suppose only one of these 'factors' is used. The plane now collapses into one of the lines in it; and the regression vector is then forced to lie along this line, which is not determined by y , but only by the x 's. There is nothing to prevent, or even to make improbable, a great increase in the angle of the new regression vector with y , with a corresponding decrease of the correlation and of the accuracy of prediction. It is possible for even a high correlation to be reduced to zero when the set of predictors is replaced by a reduced number of linear functions of them, such as might be obtained by factor analysis, and in virtually every case there will be some decrease of correlation and of the accuracy of the prediction formula obtained.

The successful determination of exact probability distributions for the multiple correlation, canonical correlations, the generalized analysis of variance and Student's t ratio, and certain other modern methods has in each case hinged on the invariance property of the statistics used. It is apparently only by restricting the choice of statistics to such invariants that the appalling number of nuisance parameters in the distributions of functions of numerous correlated variates can be reduced sufficiently to bring the sampling errors under control.

The necessity of sampling distributions in drawing any reliable inferences thus points to the use only of functions of observations possessing invariance properties, since no exact sampling distributions seem to be obtainable otherwise. It is true that asymptotic standard errors are used in default of exact distributions; but in the absence of any known bound for the error resulting from this practice, it cannot be recommended excepting in desperate extremities. Far better is the use of statistics for which exact distributions can be determined, 'exact' in the rather technical sense of freedom from any unnecessary dependence on nuisance parameters, and 'exact' also in the ordinary sense of accurate computability.

In a matrix of covariances the only thing invariant under all non-singular linear transformations of the variates is the rank. For covariance matrices of observations, constituting a sample from a multivariate normal or any other continuous distribution, the actual rank is virtually always equal to the number

of variates, and therefore has no real interest. In order to achieve invariant statistics of real value in this connection we need either *two* matrices relevant to our problem, or some relevant restriction of the class of transformations.

In the latter category is a restriction to orthogonal transformations. Under this subclass of all linear transformations there are invariants, the latent roots and latent vectors of the matrix, corresponding geometrically to the principal axes of an ellipsoid determined by the matrix. This is one way in which we are led to the method of principal components. There are also other ways, including several minimizing or maximizing properties. But one may wonder why orthogonal transformations should be given such a position of pre-eminence. A change of units of measure, for example, is a non-orthogonal transformation, but may often be desirable, and is actually made when the standard deviation of each variate is taken as its unit of measure, thus replacing the covariances by correlations.

When we come to the comparison of two covariance matrices, the number of invariants under all non-singular linear transformations is significantly increased. For a pair of symmetric square p -rowed matrices A and B , of which B at least is non-singular, a complete set of invariants is provided by the roots $\lambda_1, \dots, \lambda_p$ of the determinantal equation

$$|A - \lambda B| = 0,$$

each with an associated vector. No restriction to orthogonal transformations is now needed to get meaningful invariants.

If A and B are independent sample covariance matrices corresponding to the same population covariance matrix, the joint distribution of $\lambda_1, \dots, \lambda_p$ involves only these statistics themselves, the number of variates p , and the numbers of degrees of freedom of the two sets of sample covariances. This distribution was determined virtually simultaneously by S. N. Roy [14], [15], P. L. Hsu [12], and R. A. Fisher [7]. To test the hypothesis that two independent sample covariance matrices for the same p variates have the same expectation, one may select some function of $\lambda_1, \dots, \lambda_p$ that will be sensitive to a particular type of deviation from equality against which the null hypothesis is to be tested. There is then the problem of deriving the distribution of the particular function chosen from the Roy-Hsu-Fisher distribution, as by transformation and integration to pass from joint marginal distribution. This process has been carried out exactly for several cases for which $p = 2$, and approximately in other cases.

It is in this way that the several forms of multivariate analysis of variance may be regarded. The matrices A and B may be obtained from independent samples, or where there is replication, from the same sample, by obvious extension of the separation in univariate analysis of variance of the error sum of squares from the rest. The familiar univariate ratio F of the sum of squares of main concern to the error sum of squares, multiplied by a ratio of numbers of degrees of freedom, is generalized, if B is the matrix of error covariance estimates, by either AB^{-1} or $B^{-1}A$. Each of these matrices must be represented

by one number in order to carry out a test of significance to see whether A represents deviations substantially beyond what can be expected by chance of the sort revealed by the discrepancies between replications summarized in B . This number may be the determinant

$$|AB^{-1}| = |B^{-1}A| = |A|/|B| = \lambda_1, \lambda_2, \dots, \lambda_p,$$

whose distribution was studied, with this use in view, by S. S. Wilks [17]. It is the statistic indicated by the likelihood ratio criterion. The use of the greatest of the roots has been developed by M. S. Bartlett [1] and later papers [2, 3], and S. N. Roy in the multivariate analysis of variance [16], and presents advantages in numerous applications.

Another method makes use of the sum of the roots. This arose originally from use of the generalized Student ratio T^2 to measure the deviation of a bomb from its target, against a background of recorded errors in range and deflection of such bombs. I have written about this distribution in three symposium volumes: uses of it in *Techniques of Statistical Analysis* (1947), its mathematics in the *Proceedings of the Second Berkeley Symposium* [10], and its relation to other multivariate methods in *Statistics and Mathematics in Biology* [11]. The joint distribution of the squares of the canonical correlations is closely related to that of $\lambda_1, \dots, \lambda_p$, and has been studied by P. L. Hsu in *Biometrika* for 1941.

Problems of discrimination and classification among populations and individuals have been considered by many writers. Factor analyses appear to have little place in this problem, though it is occasionally mentioned. A. Wald proposed (*Annals of Mathematical Statistics*, 1944) a statistic for measuring the extent of misclassification, and with the help of its invariance properties reduced its distribution to a triple integral. This was studied further by H. L. Harter (*Ann. Math. Stat.*, 1951).

IV. THE DIMENSIONALITY OF A MULTIVARIATE NORMAL DISTRIBUTION

Let us consider one further problem of long standing that helps to bring out the relations between factor analysis and methods of multivariate analysis of recent development. This is the question of the dimensionality of a multivariate normal distribution. The dimensionality is the same as the rank of the matrix of the observations, the covariance matrix, and the correlation matrix. Frequently asked questions are how, on the basis of a sample, to determine this rank, and how accurate the determination is. The answers are, in one sense, easy. The rank of a sample cannot be greater than that of the population from which it is drawn. For example, if a distribution of points in three-dimensional space is confined to a plane, then every sample of these points lies in the same plane. On the other hand, for a distribution having continuous positive density of dimensionality p , the probability that a sample of more than p points should lie entirely within a flat space of dimensionality less than p is zero. If the sample consists of p or fewer points, the dimensionality is one less than the sample size. Thus, apart from possibilities whose total probability is zero, we may say that

if there are more individuals than variates, the rank of the sample matrices is exactly that of the population. This is a complete answer to the questions as they are most often asked.

This answer is however unlikely to satisfy our questioner. The rank of an empirical matrix of observations (or of their correlations or covariances) is practically always equal to the number of variates; and he wants to represent these by a smaller number of variates, and is looking for a statistician's blessing on this labour-saving procedure. His real problem is not whether in reality there are fewer dimensions than variates, but whether he will make much of an error if he proceeds as if there were. It is of course impossible to give him satisfactory assurances without knowing exactly what it is that he is ultimately going to do with his calculations, and even if one does know this it is difficult. There is a chance of considerable mischief ultimately developing from neglect of something that is really there, but since so much has to be neglected, it is not seriously out of character to suggest, for example, the method of principal components to pick out a small number of 'factors' to represent the whole lot. This can be done without any attempt to judge the 'reality' of the factors.

Another interpretation is worth noticing. Perhaps each observation is the sum of a real part and a random error, and the real parts have fewer than p dimensions, though the observations are of rank p . This is a self-consistent hypothesis. To test its correctness we need one of two things. Either we must know the matrix of covariance of the errors, or we must be able to estimate this matrix from observations. The latter requires replication, so that independent estimates shall be available of the same true part. If q independent estimates, subject to unbiased errors having the same covariance matrix, are available for each individual, a multivariate analysis of variance may be employed to test the significance of the deviations from any particular hypothetical dimensionality. This procedure is capable of providing definite tests in terms of probability, and does not require a factor analysis.

Factor analysis has often been thought of as a means of finding genetic factors; indeed, this seems to be the origin of the word 'factor' in its name. It would be more in keeping with geometrical and physical traditions to speak of 'components' of a tensor, which would be 'principal components' if possessed of certain extremal properties. The determination of genetic factors requires comparison of individuals carefully classified according to their relationships. A matrix of correlations referring to persons of unknown degrees of relationship cannot possibly prove or disprove anything about genetic factors, whatever mathematical procedures may be applied to the matrix. Thus the scope of valid applications of factor analysis in the usual sense must be reduced by the elimination of genetics.

ACKNOWLEDGMENTS

The substance of the foregoing paper was discussed at the recent International Colloquium on Factor Analysis, and a French translation published in full in the Report of the Proceedings (Paris, 1955). The author is indebted to the authorities

of the *Centre National de la Recherche Scientifique* for permission to publish an English version here.

REFERENCES

- [1] BARTLETT, M. S. (1934). Probability and chance in the theory of statistics. *Proc. roy. Soc. A*, CXXI, 518.
- [2] BARTLETT, M. S. (1948). Internal and external factor analysis. *Brit. J. stat. Psychol.*, I, 73.
- [3] BARTLETT, M. S. (1950). Tests of significance in factor analysis. *Brit. J. stat. Psychol.*, III, 77.
- [4] BOSE, R. C., and ROY, S. N. (1938). The distribution of the studentised D^2 -statistic. *Sankyā*, IV, 19.
- [5] FISHER, R. A. (1921). On the mathematical foundation of theoretical statistics. *Phil. Trans. roy. Soc. A*, CXXII, 309.
- [6] FISHER, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. roy. Soc. A*, CXXI, 654.
- [7] FISHER, R. A. (1939). The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, IX, 238.
- [8] HOTELLING, H. (1931). The generalization of 'Student's' ratio. *Ann. math. Statist.*, II, 360.
- [9] HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika*, XXVIII, 321.
- [10] HOTELLING, H. (1950). *Ap. Proceedings of the Second Berkeley Symposium*. University California Press.
- [11] HOTELLING, H. (1954). *Ap. Statistics and Mathematics in Biology*. Iowa State College Press.
- [12] HSU, P. L. (1939). On the distribution of roots of certain determinantal equations. *Ann. Eugen.*, IX, 250.
- [13] PEARSON, KARL (1926). On the coefficient of racial likeness. *Biometrika*, XVIII, 105.
- [14] ROY, S. N. (1939 a). A note on the distribution of the studentised D^2 -statistic. *Sankyā*, IV, 373.
- [15] ROY, S. N. (1939 b). p -statistics, or some generalisations on analysis of variance appropriate to multivariate problems. *Sankyā*, IV, 381.
- [16] ROY, S. N. (1942). Analysis of variance for multivariate normal populations. *Sankyā*, VI, 35.
- [17] WILKS, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* XXIV, 471.

THE DISTINCTION BETWEEN COMMON AND SPECIFIC VARIANCE IN FACTOR THEORY

By CHARLES WRIGLEY¹
University of California

Abstract. The purpose of this paper is to indicate certain inadequacies in the current use of 'communalities' in factor theory. There are considerable divergences in the way the term is defined; and several reasons for believing that, as ordinarily computed, the values become highly unstable when the number of tests or the size of the sample is increased. It is accordingly suggested that a better measure of the common factor variance might be furnished by substituting the square of the multiple correlation.

I. THE COMMON FACTOR THEORY

Factor theory assumes the variance of any test to be divisible into three parts: (i) variance which is common to other tests as well; (ii) variance which is reliable but specific to that test; (iii) error variance which arises from test unreliability. Specific and error variances may be combined to form the unique variance. This threefold distinction was introduced because of the unusual limitations shown by most psychological measurements. The psychologist naturally wants to exclude from his analysis error variance of the kind revealed by changes in scores from one occasion to another, or from parallel test forms used on the same occasion; and the reliability coefficient might seem the appropriate value to insert in the diagonal in order to achieve this. But he also seeks to exclude specific variance from his analysis, since, unless reliable variance can be shown to be present in two or more tests, it may conceivably represent something peculiar to one particular test only. These two contributions to the total variance—the specific and error variance—not only introduce further factors of no psychological interest, but may even become confounded with the common variance, thereby confusing the description of factors having psychological relevance. Hence it is commonly argued that the factorist should restrict his analysis to the common factor variance.

The term 'communality' has been defined by various writers in different and sometimes conflicting ways. Through all their definitions, however, there runs the notion that communalities are reduced values to be inserted in the leading diagonal, and that, after extraction of k common factors ($k < p$, the number of tests), they become either zero or very small. The communalities are taken to represent the 'common variance', and the variance which is thus excluded from the diagonal entry to represent the 'uniqueness'.

¹This research was supported by the United States Air Force under Contract No. AF 41 (657)—76, monitored by Air Force Personnel and Training Research Center, ATTN: Dr. John M. Leiman, Personnel Research Laboratory, Post Office Box 1557, Lackland Air Force Base, San Antonio, Texas.

The disadvantages of introducing these reduced entries are well known. Adoption of a factorial structure with more factors than tests complicates the algebraic exposition and the difficulties of calculation. There is no generally accepted method for finding the communalities; and all take considerable time for a computer with only a desk machine. Yet most factorists regard these difficulties as a reasonable price to pay for reducing the lack of precision.

In factor analysis there is always a twofold sampling. A sample of persons is selected from a population of persons and a sample of tests from a population of tests. To avoid confusion, the population of tests will be called the *domain* (Tryon's convenient term [30], p. 4), and the sample of tests the *selection*. Alternatively, Guttman's term *universe* ([15], p. 282) might be used for the population of tests, and *battery* could be used for the sample. The 'squared multiple correlation' between each test and the remaining ($p-1$) tests will be referred to as s.m.c.

II. INADEQUACIES OF THE COMMON FACTOR THEORY

The theory of common factors has four main inadequacies: (i) the existence of alternative sets of communalities satisfying the theory; (ii) the relativity of the distinction between common and specific variance, which depends in part on the particular selection of tests; (iii) the failure to distinguish explicitly between common and significant factors; (iv) when significant factors are insisted upon, the possibility of changes in the probability level adopted for significance or of an increase in sample-size resulting in more factors and increased values for the communalities. In certain forms of analysis there is a further difficulty, which is not a necessary consequence of the theory, but interlocks with it. Factors are usually rotated. If their number is increased by increasing the size of the sample, a restructuring of rotated factors is likely, with the larger factors splitting into smaller.

(i) *Alternative Sets of Communalities*. Various sets of communalities, based on varying numbers of common factors, will satisfy the theory, i.e. provide residual correlations of exactly zero [22]. The solution with the fewest common factors is generally assumed to be the best. To those who hold that factors are only operational fictions, convenient for classifying tests, this principle of parsimony presents no special difficulty. Since the s.m.c. of any test measures only common variance, a satisfactory case can be made for accepting the set of communalities with minimal deviations from the s.m.c.'s, so long as we are content with operational definitions. But to those who consider factors to represent real entities, the matter is not so easy, since it is hard to see any strong psychological reason why the solution with fewest factors is more likely to be the 'true' representation. This difficulty becomes still more acute when various sets of communalities are possible for the same minimal number of factors: see Thomson ([26], pp. 81-82) for an example.

(ii) *Relativity of the Distinction*. The distinction between common and specific variance is relative, and depends on the particular tests selected. By

adding suitable tests, specific variance can be converted to common variance. In Thurstone's analysis of perceptual tests ([28], p. 96), the white-black flicker fusion test is almost completely specific, with a communality of only 0.12. But elsewhere in the same study (p. 23), a correlation of 0.41 of the white-black flicker fusion test is reported with the blue-yellow flicker fusion test (a test excluded from the factor analysis). Presumably a flicker fusion factor would have been isolated if tests for white-black, blue-yellow, and red-green flicker fusion had all been included; and the communality for the white-black test would then have been much higher.

On the other hand, common factor variance can be reduced to specific by excluding all but one of the tests with appreciable loadings for the common factor. Thurstone's perceptual study again provides an example. The reaction time factor (Factor C) has loadings of 0.73 for visual reaction time and 0.68 for auditory reaction times, but no other loading higher than 0.36. If either test had been excluded, Factor C would probably not have been isolated; and the communality for the remaining tests would have been much smaller.

The communality also varies according as the sub-tests are scored separately or as a single composite score. In Thurstone's study of 'primary mental abilities' [27] tests of addition, subtraction, multiplication, and division were scored independently; but other investigators (e.g. the United States Army Air Force, [9], p. 80) have incorporated these into a single arithmetical test. On *a priori* grounds the composite might be expected to possess a larger common variance than any constituent sub-test, since it is longer and has greater reliability. But in practice it generally has a smaller communality. With one numerical score instead of four, variance previously regarded as common thus becomes classified as specific.

According to the mode of selection, therefore, it seems possible for the communality to range between zero and the reliability. Factorial theory can be restated in various ways to allow for this. First, we might declare that no two tests be included in the same analysis if too similar. Two reaction tests might be included, for example, when they relate to two different sensory modalities; but two flicker fusion tests would be excluded because both are visual. This seems hardly acceptable. The difficulty of deciding what is too similar is great; and investigators will seldom agree. Thurstone seems to have changed his mind about the similarity of the two flicker fusion tests after his administration of the tests. In any case, since one aim of factor analysis is to classify like tests together, it seems undesirable to prejudge the issue by classifying them before the analysis.

Secondly, the value attained for a representative finite selection of tests might be regarded as an approximation to the value which would be obtained if the analysis had been made for the whole domain. Guttman [11], [14] has investigated the consequences of this approach. If the number of common factors stays constant, then, as the number of tests grows larger, the communality will remain unchanged, but the s.m.c. will increase. The two will coincide when the number becomes infinite. Further, the s.m.c. will converge to the

communality even if the number of common factors increases, provided that the ratio of factors to tests decreases to zero as the number of tests increases to infinity. The practical difficulty is that it seems unlikely that the number of common factors will stay constant as further tests are added, or even that the ratio will in fact decrease. The 'upper bound' for the rank of the correlation matrix, derived by Sheppard ([6], pp. 200) and Ledermann [22] from available degrees of freedom, often appears to give a reasonable estimate of the number of common factors required for an exact fit. Here, however, the ratio of common factors to tests increases to one as the number of tests increases to infinity. And even this 'upper bound' may at times underestimate the number of factors. Guttman [17] provides one example, and the well-known Heywood case ([26], pp. 51-53) another. In a domain of tests, all variance hitherto regarded as specific might be expected to be converted to common, so that in the limit the communality would become equal to the parallel-form reliability.

Thirdly, we might make a virtue of necessity by stating our objective as that of designing analyses so as to convert as much specific variance to common variance as we can, thus narrowing the gap between communality and reliability so far as possible. To do so might be held to demonstrate our control of the testing situation and our power to express factorial results more generally. Or, fourthly, we might accept the relativity of common and specific variance on the ground that our concern is with the 'here and now'; in that case it would no longer be meaningful to refer to a test's communality as such; different communalities would be allowed in relation to different batteries. These last two alternatives, which accept a communality as a variable measure, probably represent the empirical situation accurately enough, and might prove acceptable operationally; but they destroy the notion of the reliable test variance as divisible once and for all into common and specific portions.

(iii) *Failure to Distinguish between Common and Significant Factors.* While the usual form of the theory allows for error introduced by unreliability, none is made for error arising from sampling of subjects. Thurstone never makes any clear distinction between a common factor, in the algebraic sense, and a significant factor, in the statistical sense. At times he implies that all error relates to self-correlations, and that the problem is solved once the reduced diagonal entries are known. But a common factor is not necessarily a significant factor. Suppose population values for all correlations are zero: there will still be some non-zero correlations in any sample, so that there will be at least one common factor. This will represent only sampling error and will not be significant. Nor is a significant factor necessarily a common factor in the Thurstonian sense. A principal axes analysis with unities in the diagonal also makes no distinction between common and specific variance: there are as many factors as tests; and, since later factors tend to have one or two loadings larger than the rest, they hardly conform to Thurstone's definition. The common variance is then so confounded with specific and error variance that such factors might be better labelled 'near-specifics'; yet they may still be significant: i.e. we may expect

similar near-specific factors on analysing another sample of the same size from the same population. Hotelling [18] and Bartlett [4] have devised significance tests for this situation.

When the problem is formulated in a purely algebraic sense such as that of reducing the correlation matrix to minimal rank, later factors may measure no more than sampling errors; yet this error will be included in the communalities and regarded as common factor variance. Our concern is therefore not with common factors *per se*, but only with those common factors that can be established as significant. As Burt has argued, "instead of aiming at a *maximum* number of common factors, a safer policy is to aim at a *minimum*, retaining merely those that are fully significant—one general factor only, if that will suffice to account for the figures observed, adding others only when the evidence compels us" ([8], p. 112).

Thurstone seeks to restate the problem in statistical terms by writing: "A matrix of minimal rank is to be obtained which deviates only slightly from the given side correlations" ([28], p. 283). But several difficulties arise from his formulation. First, it is doubtful whether the algebra of rank is intended as a guide to the number of factors, or whether his remark in the preceding paragraph that "the smallest number of independent factors that will account for the intercorrelations of n tests is the minimal rank of the correlation matrix" is meant to relate to the practical question. Second, Thurstone in his book, nowhere indicates the size of discrepancy which is small enough to be ignored.

(iv) *Dependence on the Sample of Persons and the Probability Level.* The problem of determining the number of significant common factors has been considered by Lawley [21] and Rao [25]. Increase in sample size usually results in reduced standard errors and in the attribution of significance (as the accepted probability level) to variations which would otherwise have been taken to represent merely sampling effects. It might be expected, therefore, that the number of factors would increase (a) if a more lenient level were adopted, or (b) if the sample of persons were increased. This in turn would lead us to identify as common some of the variance hitherto accepted as unique.

Is there any evidence that this situation fails to hold for common factors? Possibly the Thurstonian hypothesis of a limited number of common factors might be restated to apply to a population of persons, but (owing to random errors) not to a sample. As sample size increased, exact reduction of rank would be expected with fewer factors, instead of with more, because factors which resulted only from sampling error would grow smaller, and would disappear if correlations could be obtained for the population. On this view, increasing the size of sample would be offset by smaller residuals, and the number of significant factors would remain unchanged.

Two comments may be made. First, present evidence, though inconclusive, seems opposed to this view. Current neurological and genetic knowledge suggests greater complexity in the determination of the behaviour of the organism than can be accounted for by any faculty-type theory of a limited number of primary factors. With larger samples of persons and wider selections of tests

factorial results have become more and more complex. The theory of a restricted number of common factors might be tested empirically by finding whether the size of later factors diminishes as the size of the sample increases (implying a better fit to the common-factor hypothesis as sampling error diminishes). Secondly, this restatement requires that sample size be large enough to identify the full set of factors, but does not indicate how large a sample is needed. As we increase our sample, the number of common factors and the estimates of the communalities would change until the unknown number of common factors was reached.

As long as no concrete evidence is available in its support, we should perhaps view any hypothesis of a limited number of common factors with scepticism. It seems safer to regard the number of significant common factors as depending on sample size, with its corollary that the communalities will vary from one sample to another. If so, the assumption of a fixed division of reliable variance into common and specific portions again becomes untenable. As pointed out earlier, this relativity could be accepted by restating the objective as that of trying to narrow the gap between communality and reliability; in that case a large specific variance would be taken as indicating poor experimental design. A bigger sample might therefore be required to ensure (a) better determination of common variance, (b) more precise estimation of loadings, (c) isolation of small common factors otherwise unrecognised.

The most acute problem arises, however, because increase in the number of factors not only provides us with the additional information represented therein, but also modifies the structure of the remainder when factors are rotated. Suppose for the moment the objective to be finding diagonal values that yield residuals with the smallest sum of squares for an agreed number of factors. First, any increase in the number of factors may lead to (a) radical changes in diagonal values (they sometimes decrease when the number of factors increases), and (b) considerable changes in the factor loadings, with alterations in signs as well as the size, as Burt has pointed out ([8], p. 114). This source of restructuring is peculiar to the communalities formulation because their values are not known; the problem could be overcome by using some fixed value, e.g., s.m.c.'s, reliabilities, or unities, in the diagonal.

Secondly, when factors are rotated, increasing their number causes the larger factors to split into smaller: this happens even with physical measurements, e.g. when the quartimax procedure [24] is applied to the Pearson matrix analysed by Burt [8]. Unities in the diagonal may have increased this tendency; but Guilford ([10], p. 286) has also noticed it. No doubt it is responsible for some of the disagreement in the literature dealing with certain factors. The problem needs empirical investigation; but three observations may here be made.

1. To begin with, the problem is not confined to communalities. When the number of factors is made to depend on a test of significance, it becomes in part a function of size of sample, no matter what the diagonal entries may be. Indeed, changes in structure are likely to be less for communalities than for

unities or reliabilities (though greater than with s.m.c.'s), because reducing the diagonal values reduces the smaller factors proportionately.

2. Before including factors in a rotation, we might require them to be not only significant but also large enough to be psychologically important. Psychologists should certainly not rotate more factors than are significant; but it might be advantageous to operate with fewer. The map-maker does not try to put in all the known detail; and similarly the factorist might find loss of detailed information counterbalanced by increased clarity. Reduction might be achieved in one of two ways:

(a) We may introduce a rule for excluding factors which, though significant, are not of psychological importance. Kaiser [19] has proposed the exclusion of all factors for which the sums of the squares of the loadings are less than 1.00. It seems a doubtful gain to accept in the system a factor which contributes less information than any single test which it replaces. Tryon ([31], p. 17) suggests that the practical importance of a factor might be better judged by its highest loading than by its overall contribution; in that case we might reject every factor whose loadings were all under (say) 0.15.

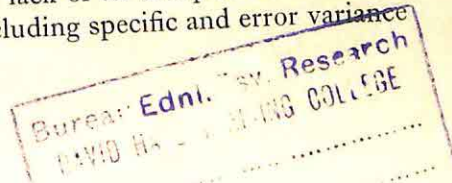
(b) We may adopt that particular set of rotational solutions which achieves greatest clarity. This would require a definition of clarity and a way of measuring it: e.g. it might be argued that the clearest structure was achieved when the variance of the squared loadings in the rotated solution was greatest, the quartimax principle being used to decide the number of factors as well as the preferred rotation. Increasing the number of significant factors would then make no difference beyond the point where greatest clarity was attained.

3. The problem is greatly reduced if a hierarchical structure is adopted. Possibly because its basis lies in faculty psychology, simple structure theory has tended to regard each factor as coordinate in status with every other. Burt [7], followed by Vernon [32], considers a hierarchical organization to be far more effective than a coordinate one for classifying tests. This can be illustrated from the analysis of ability. If only one factor is extracted, a general factor of intelligence will probably be obtained. If two factors are extracted, verbal tests will probably be separated from spatial and mechanical. If further factors are extracted, the verbal and the mechanical factors will each subdivide. Thurstone's recent advocacy of second-order factors ([20], ch. 18) seems to represent a similar step towards hierarchical factors; so does Guilford's schematization of factors for thinking and memory [10].

A hierarchical system is akin to a genealogical chart: large factors are progressively subdivided into smaller. When the sample is small, few factors are significant, and only the higher levels can be established; with a larger sample, the chart can be filled in with greater detail. The rotation of greatest clarity is then represented by a horizontal segment across the chart.

III. LACK OF AGREEMENT IN THE DEFINITION OF COMMUNALITY

One reason for all these difficulties is the lack of an accepted definition of communality. Most factorists agree about excluding specific and error variance



and invoking as few common factors as possible; but they differ as to how these aims are to be realized. Some, like Spearman, Thurstone, and Guttman, start with a hypothesis about the formal structure of the tests, and define the communalities as the values ensuring the best fit to the postulated model. Others, e.g. Tryon, and Guttman in some of his contributions, define the communality in terms of relations between a particular test and a domain. Others again, like Lawley and Rao, define the communality in statistical instead of algebraic terms².

Each of these alternatives deserves a brief examination. Until there is some agreement on the definition of communality, we can scarcely hope for any agreement about the values to be accepted.

The Spearman Model. Spearman assumed that, once specific and error variances had been removed, a single factor would suffice to describe the correlations. He himself scrupulously refrained from inserting any figures in the diagonal of his correlation tables. But, if his tables are to be factorized in the usual way, then the adoption of his theory would entail choosing for the communalities those values which most nearly reduce the matrix to rank one. To demonstrate the validity of the theory, it is necessary to test it with a wide variety of correlation matrices. This test, however, clearly revealed the need for rejecting the Spearman hypothesis: the majority of the correlation matrices actually obtained could not be reduced to rank one. Evidently the 'two factor theory' was far too simple to accord with the facts.

²Pearson's method of analysis into 'principal axes' assumes that the diagonal values will each be unity [8]. The first to substitute reduced values seems to have been Burt. In his earliest papers (*Brit. J. Psychol.*, III, 1909, pp. 94f. etc.) the tables of observed correlations have reliability coefficients (within brackets) inserted in the diagonal cells; but to reconstruct the theoretical correlations to get the observed correlations he uses the formula $r_{ij} = r_{ig} r_{jg}$: so that $r_{ii} = r_{ig}^2$. The saturations themselves (r_{ig} , r_{jg} , etc.) are based on the sums of each column. In later papers he explains more fully the method of summation adopted to calculate the saturations. Where the reliability coefficients include a significant contribution from the specific factor (as in fact is most commonly the case), "a slight but obvious complication" (he says) "is necessary in applying formula iii" (the summation equation for the saturation coefficient), i.e., as the context indicates, it is necessary to reduce the reliability coefficient to a value agreeing more closely with the general trend [5, p. 53, footnote].

Thomson, quoting this and the corresponding passages in *Factors of the Mind* (pp. 448, 462) thinks it is still "not quite clear how he (Burt) filled in the blank diagonal cells" (26, p. 25). Burt, however, maintains that the use of a fixed rule (as with Thurstone's suggestion of using the highest correlation in the column) tends to yield misleading results. Since his own tables were always arranged as nearly as possible in hierarchical order, a process of 'smoothing' seemed the quickest and most accurate way to obtain a provisional value; and the value so obtained was adjusted (when necessary) by successive approximation. However, when the hypothesis to be tested was the adequacy of a single factor only, it was possible to modify the summation formula so as to omit the values for the 'self-correlations' altogether. This proved desirable in discussions with Spearman, since he held that the method of successive approximation was "arguing in a vicious circle". Unfortunately when more than one factor was assumed, no such device was available. Sheppard, however, pointed out that the single factor hypothesis was equivalent to postulating a matrix of rank one, and that, "when more than a single factor was present, the appropriate value would be the sum of the squares of all the saturation coefficients"; and this was adopted as a convenient check in most of Burt's later papers (Burt [8] and personal communication).

The Thurstone Model. At present Thurstone's hypothesis is probably the most widely accepted. Reverting to faculty psychology, he postulated a limited number of 'primary mental abilities'. As we have seen, his theory was stated algebraically as that of attaining minimal rank for the correlation matrix³. However, the definition of communality in terms of rank has proved difficult arithmetically. For one thing, no easy method has been developed for finding minimal rank: Albert's procedure [1], [2] is too complex for regular use. Moreover both the number of factors and the values of the communalities have to be determined; and this double determination is all the harder because the two estimations are interdependent. Furthermore, if minimal rank be required in the literal sense of a solution that reduces all residuals to zero, then it would seem that the number of common factors can rarely be reduced below Lederman's 'upper bound', and sometimes not even as low as that. Hence in practice far more factors have to be accepted than Thurstone apparently anticipated when formulating his theory. For example, when s.m.c.'s are inserted in the leading diagonal, the well-known Holzinger and Harman correlation matrix, based on 24 ability tests, has 13 positive latent roots: this means that at least 13 factors are needed to account for the correlations ([15], p. 153).

What is even more important, the actual factorial findings seem to be making Thurstone's hypothesis as unacceptable as Spearman's. Thurstone never stated explicitly how many factors in all might be expected. His adoption of the phrase 'primary mental abilities' in his analysis of cognitive tests [27] suggests that he expected a rather limited number; and his repeated insistence on the importance of parsimony supports this interpretation. Yet the number of factors empirically identified continues to increase. Guilford [10] lists no fewer than 40 in his review of work on thinking and memory alone, and expects to discover more when appropriate tests are constructed.

The Guttman Models. Guttman ([13], ch. 6) notes the inability of Thurstone's theory to fit the data, and has proposed new formal models such as the simplex, the circumplex, and the radex. These employ different relations among factors loadings from those ordinarily adopted (viz. accounting for correlations by the products of loadings). He accordingly defines communalities as the values most nearly maintaining the postulated structure. Up to the present, however, no allowance seems to have been made for errors resulting from sampling. As with Spearman and Thurstone, therefore, we can only accept Guttman's definition if his postulated structures are shown to be applicable to a wide variety of empirical matrices.

³Thurstone's account [29, pp. 282-283] is not altogether clear. Thus in a footnote on p. 283, he expresses his objective as finding the values which maintain the rank of the off-diagonal elements in the correlation matrix. This may have been to avoid certain criticisms from Lederman [22], who had demonstrated that anomalous correlation matrices might occur in which minimal rank was not associated with a minimal sum of communalities. It is difficult to give precise meaning to Thurstone's remarks about the rank of the off-diagonal entries unless we assume a minimal rank requirement. For similar criticisms of this passage the reader may refer to this *Journal*, IV, p. 144, and VII, p. 108.

The Tryon Model. Tryon [20] formulates the problem quite differently. He defines the communality in terms of his domain-sampling theory, as the squared correlation between the test and a domain of tests whose correlations with the remaining $p-1$ tests are proportional. With this definition, the communality of a test can be shown to equal the correlation between it and a hypothetical test having identical correlations with the remaining $p-1$ tests in the selection. Correlations are then estimated between the tests and their hypothesized counterparts by psychometric formulae of the Spearman-Brown type. He thereby obtains a set of quadratic equations which when solved give exact values for the communalities ([30], p. 246)⁴.

The procedure has been restated by Kaiser [20] in determinantal form; and the resultant iterative technique is found to converge for artificially constructed matrices. It is not yet known whether with empirical matrices convergence is slow or fails to occur. Mathematical properties of the method have also been examined by Guttman [15]. The conception of a test as drawn from a hypothetical domain of similar tests is attractive, since psychometric theory has usually paid more attention to the sampling of persons than to the selecting of tests. The power of the method is shown by the fact that it leads to explicit equations, and that their solution requires no knowledge or estimate of the number of factors.

On the other hand, to postulate a domain of tests with proportional profiles is only a necessary, not a sufficient, condition. There are further requirements unstated. For instance, parallel forms of a test, as encountered in reliability theory, would have proportional profiles, but would not meet the communality requirements, because drawn from the same behaviour domains rather than from different. The distinction between communality and reliability, so Tryon would hold, is that the domain for the reliability must be defined in advance: items drawn randomly from this predesignated domain would provide the requisite parallel tests; but the domains for the communality are operational fictions defined after the tests have been selected for analysis.

As we have seen, alternative sets of communalities are possible; and it is by no means clear why Tryon's procedure should lead to the minimal rank set, though it certainly seems to do so, at least with artificially constructed matrices. Possibly this follows from the particular form which the linear dependence of each common-factor portion on all the others assumes. Because of this assumption of rank reduction, the practical consequences of the Tryon-Kaiser method seem to be mostly similar to Thurstone's, and thus subject to the same criticisms. First, no allowance is made for errors arising from the sampling of persons.

⁴Using the same concepts, Tryon has developed several different methods for determining the communalities: see [30] and [31]. Kaiser derives his determinantal equation in terms of Thurstone's common-plus-unique partitioning rather than of Tryon's correlation with a hypothetical test having an identical correlational profile. He argues that the rank approach of Thurstone and the domain of Tryon lead to the same arithmetic and the same practical consequences. In his view, although a formulation in terms of rank leads to the simpler algebra, the logic based on the 'domain' concept yields a fuller insight from the standpoint of psychology.

An algebraic solution is designated in which all residuals are reduced exactly to zero. Hence a large number of factors will generally be needed in any small sample—more perhaps than might reasonably be expected to be meaningful. Secondly, Tryon explicitly admits that communalities will change as extra tests are added. 'Proportional tests' are merely fictional devices, introduced in order to achieve a more compelling conceptualization; it is not asserted that such tests could be developed empirically. But definition of the communality in such terms means that the profile has to be extended to cover further tests added to the selection.

For Tryon, therefore, the communality is always restricted to the context of the tests in a particular battery. Within that context he considers that, except for sampling variations, the communality will have only one value. Since he acknowledges that diagonal entries should be based on the 'here and now', i.e. on the particular selection of tests chosen by the investigator, one wonders whether a case might not be made for using the actual s.m.c. which measures variance held in common with the remaining $p-1$ tests, instead of a fictional s.m.c. This will be discussed later.

The Statistical Model. The final position to be considered is that in which the problem is stated in statistical rather than algebraic terms. This is represented by Burt's procedure of correcting for uniqueness ([8], p. 113), Lawley's maximum likelihood method [21], and Rao's canonical method [25]. The communalities are then the values which enable the correlations (weighted in accordance with the communalities) to be fitted as well as possible for the significant factors, i.e. they are those which give the smallest sum of squares of weighted residuals.

This includes some Thurstonian features, but rejects others. The number of factors depends on significance, not on rank; but rank reappears in the need to find diagonal values providing the best fit between correlations and factor, i.e. which achieve rank reduction, not to the minimum, but to that implied by the indicated number of significant factors; and these diagonal values are used in applying significance tests. Many disadvantages of Thurstone's algebraic approach are overcome by such methods. The requirement that common factors should also be significant factors is obviously in line with modern statistical theory. Furthermore, the method is free from some of the theoretical difficulties of Thurstone's position. The distinction between common and unique variance is no longer an attribute of a test *per se*, a position hardly tenable in view of variability of the communality: instead it is made relative to the particular set of tests and persons, and its justification resides simply in the fact that the loadings provide the best fit to the correlations. We are no longer committed either to the faculty theory of a limited number of factors or to the postulate that their number either stays constant when tests increase, or increases at a slower rate.

Two difficulties remain. The first is logical. While dependence of the communality upon the test selection may be psychologically acceptable, it is hard to give psychological meaning to the variations in the communality that arise

from changes in the sample of persons. The approach seems to imply that in a very large sample practically all common factors will be significant, so that the values for the communalities would then approach those that would be secured by minimizing rank. The maximum likelihood communality is therefore a biased statistic. Samples nearly always give too low a value, because communalities are estimated, and factors are tested for significance, in a single calculation. The two problems might be separated by first determining communalities algebraically, e.g. by taking the minimal rank values or values conforming to Tryon's definition. These might be regarded as estimates for the selection of tests for a population of persons. There would remain an error of estimate depending on sample size; but there would be no inevitable bias on the low side. Having calculated the factors derived from these communalities, we could then test the hypothesis that any particular factor is not different from zero. The estimational procedure would thus be separated from the test of significance.

In Rao's procedure there is a second weakness. For the computer the iterations for determining the number of significant factors and the best diagonal entries are very laborious, and in general are practicable only with an electronic computer. This may not prove a permanent difficulty. Calculations might perhaps be accelerated by better initial estimation and by extrapolation to speed convergence. Moreover, faster electronic computers may soon be available.

IV. COMMUNALITY, RELIABILITY, AND SQUARED MULTIPLE CORRELATION

The central argument of this paper has been that the distinction of common variance has not been shown to be stable. This section will consider the consequences if communalities should prove relative to the selection of tests and persons.

The Communality as a Domain Measure of Common Variance. Thurstone and Tryon agree in regarding the communality as a measure of variance common to a test and its domain, but define the domain differently. Thurstone postulates a domain of tests with the same factorial composition as those already selected. He does not state this explicitly; but, as Guttman has shown, it follows from his hypothesis. Since no satisfactory evidence has so far been provided to prove that the number of factors stays constant, insistence on an unchanging boundary between common and specific variance seems unwarranted. Can Thurstone's position be modified to meet this difficulty? The distinction between 'common' and 'specific' might be made one of degree by conceding that further common factors can be obtained by adding tests, while maintaining that, if the selection of tests was representative at the outset, the further factors will be small and of little psychological relevance. But such a restatement seems unconvincing. First, even small factors might lead to an appreciable change in a rotated solution because of additional dimensions. Secondly, since the theory specifically implies that the communality is a measure of the common variance of a domain, even the small factors should not be excluded in estimating it:

otherwise calculated values would be merely an uneasy compromise between the value for a domain and that for the particular selection.

Tryon, on the other hand, postulates a different domain for each test. Thus in a different way he requires, like Thurstone, that no new common factors be introduced in determining the common variance in the domain. The appearance of new common factors he meets by redefining the domains. Once this is conceded, it would seem to follow that, by adding a sufficiently large and representative set of tests, all specific variance will be converted to common. In the limit, therefore, the communality would equal the reliability: by reliability we here mean the correlation between a test and a parallel form, since a parallel form could be added as a further test. A thoroughgoing adoption of domain-sampling therefore seems logically to require abandoning the common-specific distinction altogether. In the domain, test variance will be divisible into two parts only, reliability and error, so that reliabilities should be inserted in the leading diagonal.

The Communality as the Value Giving the Best Fit to the Observations. An alternative case can be made for communalities on the ground that they provide the smallest residuals. Thurstone apparently hoped that communalities would give the best correlational fit, as well as measure common variance in a domain. This would be providential. But the coalescence could only be achieved by the unproven assumption of a limited number of common factors in the domain.

The correlations merely offer a summary description of the relations between the test measurements. Hence if the aim is to fit the observed data as closely as possible, it would seem better to seek to reproduce the test measurements: i.e. if there is to be a reduction from p tests to k factors, the factors should be those for which the factor measurements are most highly correlated with the test measurements. This is the approach of the maximum likelihood method. Factors so secured are those which provide the maximal canonical correlations with the tests; and Rao accordingly prefers to call the procedure the *canonical* method of factor analysis ([25], p. 94). The disadvantages of this formulation, if we may judge from certain results when using the maximum likelihood method, seem to be the possibility of obtaining diagonal values which do not conform to customary expectations: e.g. when the sample is so small that only one factor is significant, a canonical correlation between tests and factor of 1.00 can be secured by reproducing one test as the factor. More generally, if the problem is to maximize the relations between known tests and unknown factors which are functions of the tests themselves, something resembling spurious correlation may conceivably arise. This seems to hold for all communality formulations. Since the factor loading is the correlation between test and factor, the communality is equal to the correlation of any test with the full set of common factors. The test is therefore to be correlated with factors of which it forms part. Such a theory seems to leave us free to arrange the conditions of the right-hand side of the test-factor relation so that we can "pull ourselves up by our own bootstrap".

The Squared Multiple Correlation as a Measure of the Common Variance in the Test Selection. The possibility of using s.m.c.'s deserves greater consideration than it has so far received. Both in theory and in practice the communality has proved a somewhat intractable concept. If s.m.c.'s. were inserted in the leading diagonal, the distinction between common and specific variance would be based on the test selection instead of a hypothetical domain. It is well known that the s.m.c. is a lower bound for the communality. The s.m.c. measures variance common to the test and the remaining $p-1$ tests in the selection, whereas the communality measures that common to the test and the factors based upon the p tests. The additional 'common variance' introduced by using communality instead of s.m.c. is therefore a measure of the extent to which the test itself is reproduced by the factors. Can this fairly be called common variance? Common variance initially was considered to be variance common to two or more tests, but the communality is now seen to include a certain amount of variance that is common only to the test and a transformation of itself. It is scarcely psychologically meaningful to call this additional portion 'common variance'. The dilemma could have been avoided had it been shown that the increment represents extra common variance obtained by increasing the selection to the limit; but this line of argument leads to the reliability instead of to the communality.

Various objections raised against communalities can be met, or be partially met, by the use of s.m.c.'s. The problem of alternative sets of diagonal values satisfying the theory then disappears. S.m.c.'s are probably less influenced by increasing the sample of persons, since the revised values, though more accurate, may either be higher or lower. Determining the diagonal entries becomes separate from determining the number of significant factors. One disadvantage is the lack of a significance test; but this is a problem for the statistician rather than the psychologist.

The common-specific distinction is still relative to test selection, since s.m.c.'s increase when further tests are added. Hence knowing the common variance would provide no information about the test *per se*, transferable to other analyses, but would indicate the relation between that test and the other tests included in the analysis. Marked discrepancies between s.m.c.'s and parallel-form reliabilities might suggest defects in test selection. Computationally there is an attractive precision about the s.m.c.'s. On any high-speed machine they can be obtained much more easily than communalities.

Two related problems need investigation: (a) the size of the increments from s.m.c.'s to communalities; (b) the extent to which factorial results are changed by using s.m.c.'s for communalities. In default of more direct evidence, it must suffice to consider the logical alternatives and their implications. At one extreme there is the possibility that increments remain constant from variable to variable: the two sets of principal axes loadings would then be proportional, the only difference being that the latent roots are greater with the latter than with the former by the value of the constant; so that with

communalities the minor factors play a larger part in rotations. In this case therefore the use of s.m.c.'s would make little difference in practice ; but the s.m.c.'s would be simpler computationally.

The key issue is then represented by the other extreme, namely, where increments differ greatly from one variable to another. Here communalities have definite algebraic advantages, provided that they can be accurately determined : (a) the correlation matrix is then Gramian, i.e. there will be no negative latent roots with associated imaginary loadings ; (b) the diagonal values with which the analysis starts are reproduced by the factor loadings. But even here the psychological advantages seem to rest with the s.m.c.'s. Consider the artificial correlations in the table below. In the first case the communalities are approximately proportional to the s.m.c.'s ; in the second the factor loadings have been chosen so that there will be a marked discrepancy between the s.m.c. and the communality for the fifth variable. For this the correlations in the second case are only 0.045 larger than in the first ; yet we are asked to believe that

TO ILLUSTRATE THE EFFECTS OF MINOR CHANGES IN FACTORIAL
STRUCTURE ON THE COMMUNALITIES AND THE SQUARED MULTIPLE CORRELATIONS

Factor loadings		Correlations					Communalities	Squared multiple correlations
0.90	—	—	0.72	0.63	0.54	0.09	0.810	0.622
0.80	—	0.72	—	0.56	0.48	0.08	0.640	0.546
0.70	—	0.63	0.56	—	0.42	0.07	0.490	0.426
0.60	—	0.54	0.48	0.42	—	0.06	0.360	0.315
0.10	—	0.09	0.08	0.07	0.06	—	0.010	0.009
0.90	0.05	—	0.7225	0.6325	0.5425	0.1350	0.812	0.626
0.80	0.05	0.7225	—	0.5625	0.4825	0.1250	0.642	0.549
0.70	0.05	0.6325	0.5625	—	0.4225	0.1150	0.492	0.429
0.60	0.05	0.5425	0.4825	0.4225	—	0.1050	0.362	0.318
0.10	0.90	0.1350	0.1250	0.1150	0.1050	—	0.820	0.022

in the first case the common factor variance is only 1 per cent of the test's total variance, whereas in the second case it amounts to 82 per cent. If someone unfamiliar with factorial theory were confronted with these two cases, he would surely regard the s.m.c. as offering the better measure of the 'commonness' of the fifth variable. No assertions that the communality is the value attained by the s.m.c. in a domain of tests would affect his verdict. Rather he would believe that Thurstone had allowed the algebra of the situations to dominate over the psychology.

S.m.c.'s, of course, do not answer all the issues raised in this paper. Thus, if significant factors alone are rotated, their number may increase as the sample of persons increases ; and that may lead to a radical restructuring. The fact

that the factorial results may very possibly change with increase in the sample of persons, or in the selection of tests, should warn us that any hope that factors can be regarded as real causal entities is unlikely to be realized. Factors must be regarded as operational fictions. Our task is to fix the rules governing their extraction so that our practical problems, e.g. the classification of the tests, can be most skilfully and most expeditiously handled. And it certainly seems possible that s.m.c's may serve these purposes better than communalities.

V. SUMMARY

This paper considers the distinction between specific and common factor variance and the possibility of separating them by using communalities. If such a distinction is to provide information about the tests, it seems essential to demonstrate the stability of the communality from one battery of tests to another and from one sample of persons to another.

There are three principal reasons for believing that the actual values found cannot be stable. First, the existence of alternative sets of communalities in any one analysis requires some further restriction to decide between them ; and any such restriction seems to depend more on human convenience than on any fundamental test characteristic. Secondly, standard theory seems to have made inadequate allowance for sampling errors when the same test battery is factored for different-sized samples of persons. To achieve generality, factors must be significant as well as common. The number of factors and the size of the associated communalities will therefore depend on the number of persons in the sample. Thirdly, difficulties arise when further tests are added to the battery. If no new common factors were introduced as a result, the communality for any test would remain constant ; but empirical findings do not support this expectation. Illustrations have been given which show (a) how the addition of appropriate tests will convert specific to common variance, thereby increasing the communality, and (b) how removal of tests will reduce common variance to specific.

Once the distinction between common and specific variance is recognized as operational, i.e. as relative to the particular testing situation, we are led to ask whether the communality is the most effective instrument for reducing the number of factors. If our concern is with common variance in a domain, a better case could be made for reliabilities, since all the specific variance can in theory be converted to common by adding enough tests. The principal claim of communalities rests on the fact that they give a better fit than any other diagonal values to the correlations actually observed. The psychological advantages, however, appear to rest with the squared multiple correlations, which measure the proportion of common variance in any particular test selection.

REFERENCES

- [1] ALBERT, A. A. (1944). The matrices of factor analysis. *Proc. Nat. Acad. Sci.*, XXX, 90-95.
- [2] ALBERT, A. A. (1944). The minimum rank of a correlation matrix. *Proc. Nat. Acad. Sci.*, XXX, 144-146.
- [3] BANKS, C., and BURT, C. (1954). The reduced correlation matrix. *Brit. J. Stat. Psychol.*, VII, 107-118.
- [4] BARTLETT, M. S. (1950). Tests of significance in factor analysis. *Brit. J. Psychol. Stat. Sect.*, III, 77-85, IV, 1-2.
- [5] BURT, C. (1917). *Distributional and Relations of Education Ability*, London : P. S. King.
- [6] BURT, C. (1952). Bipolar factors as a cause of cyclic overlap. *Brit. J. Psychol. Stat. Sect.*, V, 197-202.
- [7] BURT, C. (1949). The structure of the mind ; a review of the results of factor analysis. *Brit. J. Educ. Psychol.*, XIX, 100-111, 176-199.
- [8] BURT, C. (1949). Alternative methods of factor analysis and their relations to Pearson's method of 'Principal Axes'. *Brit. J. Psychol. Stat. Sect.*, II, 98-121.
- [9] GUILFORD, J. P., ed. (1947). *Printed Classification Tests*. Washington, D.C. : U.S. Govt. Printing Office. Army Air Forces Aviation Psychology Program Research Report, No. 5.
- [10] GUILFORD, J. P. (1956). The structure of intellect. *Psychol. Bull.*, LIII, 267-293.
- [11] GUTTMAN, L. (1940). Multiple rectilinear prediction and the resolution into components. *Psychometrika*, V, 75-99.
- [12] GUTTMAN, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, XVIII, 277-296.
- [13] GUTTMAN, L. (1954). A new approach to factor analysis : the radex. In Paul F. Lazarsfeld, *Mathematical Thinking in the Social Sciences*. New York : Columbia University Press.
- [14] GUTTMAN, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, XIX, 149-161.
- [15] GUTTMAN, L. (1956). Best possible systematic estimates of communalities. *Psychometrika*, XXI, 273-285.
- [16] GUTTMAN, L. (1956). Improved bounds for communalities. *Research Report*.
- [17] GUTTMAN, L. (1956). To what extent can communalities reduce rank ? *Research Report*.
- [18] HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, XXIV, 417-441, 498-520.
- [19] KAISER, H. F. (1956). The varimax method of factor analysis. Ph.D. Thesis, on file in the University of California Library.
- [20] KAISER, H. F. (1956). Solution for the communalities : a preliminary report. *Research Report*.
- [21] LAWLEY, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edin.*, LX, 64-82.
- [22] LEDERMANN, W. (1937). On the rank of the reduced correlation matrix in multiple factor analysis. *Psychometrika*, II, 85-93.
- [23] LEDERMANN, W. (1940). A problem concerning matrices with variable diagonal elements. *Proc. Roy. Soc. Edin.*, LX, 1-17.
- [24] NEUHAUS, J. O., and WRIGLEY, C. F. (1954). The quartimax method : an analytic approach to orthogonal simple structure. *Brit. J. Stat. Psychol.*, VII, 81-91.
- [25] RAO, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, XX, 93-111.
- [26] THOMSON, G. H. (1951). *The Factorial Analysis of Human Ability*. London : University of London Press.

- [27] THURSTONE, L. L. (1938). *Primary Mental Abilities*. Chicago : University of Chicago Press.
- [28] THURSTONE, L. L. (1944). *A Factorial Study of Perception*. Chicago : University of Chicago Press.
- [29] THURSTONE, L. L. (1947). *Multiple Factor Analysis*. Chicago : University of Chicago Press.
- [30] TRYON, R. C. (1956). Communality of a variable : formulation by cluster analysis. *Research Report*.
- [31] TRYON, R. C. (1956). Cumulative communality cluster (CCC) analysis. *Research Report*.
- [32] VERNON, P. E. (1950). *The Structure of Human Abilities*. London : Methuen.

THE RELATIVE INFLUENCE OF HEREDITY AND ENVIRONMENT ON ASSESSMENTS OF INTELLIGENCE

By CYRIL BURT AND MARGARET HOWARD

The Application of the Parallelogram Law. In several previous publications we have briefly referred to a graphical method of representing and computing the effects of genetic and non-genetic factors which is based on the same principle as the familiar diagram used for illustrating the composition and resolution of forces in physics—the diagram popularly known as the ‘parallelogram of forces’. The method we propose treats the variables concerned as measurable tendencies or ‘vectors’ obeying the ordinary rule for the addition of vector quantities. It has proved specially helpful in attempts to analyse and interpret the factors involved in measurements of intelligence. Results obtained in this way were cited in a previous paper on multifactorial inheritance in this *Journal* (IX, pp. 126–7). Several correspondents have since asked for a more detailed description of the procedure. We therefore offer a short account of the working method, illustrated by data given in the article just quoted.

In explaining the principles on which the calculations are based, we should like to emphasize at the outset that we regard the values so reached as representing only approximate estimates for the hypothetical variables concerned. The approximation is, we believe, sufficiently exact for use with measurements of the kind we have in view, since their error or unreliability is in any case likely to be rather large; but we do not claim that the method will furnish the most efficient estimates, such as might satisfy the ideals of the mathematical statistician. Indeed, Mr. Alan Stuart, who has been good enough to read and comment on an early draft of this note, states that in his opinion the type of problem here raised is one which still awaits an adequate theoretical investigation.

Hypothetical Analysis. In applying the procedure to the problem mentioned we may start by dividing the conditions which affect our assessments of intelligence into the following hypothetical factors.

I. *Genetic or Heritable Factors (H).* These include (i) the *Fixable* contributions resulting from ‘the essential genotypes’; (ii) the *Non-fixable* contributions due chiefly to the effects of partial dominance; and in addition to these it is usually desirable to allow for (iii) the effects of *Assortative Mating*. (Formal definitions of these terms are given in the article already quoted, p. 106.) For our present purpose it will not be necessary to deal with these subclasses separately.

II. *Non-genetic or Non-heritable Factors (N).* These include (i) *Random Environmental Effects*, i.e. the effects of environment so far as it operates in complete statistical independence of the genetic tendencies (*R*), (ii) *Systematic Environmental Effects*, i.e. the effects of environment so far as it operates in the same direction as the genetic tendencies (*S*), (iii) a contribution due to the

Interaction of heredity and environment (if any), i.e. to the joint effects of both when they are correlated (*J*), (iv) all other unidentified influences—errors of measurement, sampling errors, etc., which may be summed up under the general term *Unreliability* (*U*).

Data. The data available (collected from a sample of nearly 1,000 children) consist of (a) the crude marks obtained from group tests of the usual type (verbal and non-verbal), (b) the adjusted assessments obtained by submitting the initial marks for criticism or correction to teachers and then adjusting the marks where necessary by supplementary tests of an individual and non-verbal type. In addition we have assessments, obtained from somewhat smaller groups, for the apparent intelligence of the parents, of the nearer relatives, and of twins, as well as ratings for the economic and cultural conditions of a representative sample of families.

From these various measurements or assessments we have computed the following coefficients:

(i) Correlations between two successive tests or assessments: these enable us to assess the amount of 'unreliability';

(ii) Correlations between identical twins who have been (a) brought up together and (b) separated from infancy and brought up apart: these enable us to estimate the effects of post-natal environment (intra-twin variance).

(iii) Correlations between offspring of first cousins: these by exclusion should allow us to estimate the possible effects of pre-natal environment; in point of fact, however, the effects appear so small as to be virtually negligible.

(iv) Correlations between (a) parents and offspring, (b) siblings, (c) fathers and mothers: these enable us to assess the relative influence of the first three 'heritable' conditions separately (if required) as compared with the influence of environment.

Preliminary Analysis of Variance. Using these correlations, and relying mainly on the formulae derived in our previous paper, we can partition the total variance for each set of assessments as follows (Table I: for the detailed working see *loc. cit.*, IX, pp. 124f).

TABLE I. ANALYSIS OF VARIANCE

	Crude Test Marks	Adjusted Assessments
Heredity	77.06	87.56
Environment: random effects	5.91	5.77
Environment: systematic effects	10.60	1.43
Reduced Total	93.57	94.76
Unreliability	6.43	5.24
Total	100.00	100.00

Analysis by Vectors. We now propose to regard the main components as vector quantities (i.e. quantities defined solely by their magnitude and direction)

and to treat the observed assessments as a resultant obtained by combining the components in accordance with the 'parallelogram law' for the addition of vectors, on the analogy of the resultant and component forces in mechanics.

In the problem we are examining here we are not concerned with the effects of unreliability. These form a random variable uncorrelated with the remaining tendencies, and would therefore have to be represented in a third dimension. Here we shall be concerned solely with an attempt to partition what we may call the 'reduced total', i.e. the total left after the effects of unreliability have been deducted. Accordingly, on the basis of the first three percentages shown in Table I we may now re-analyse the data in the following way.

So far as direct causation is concerned the variations in a child's genetic constitution are completely independent of the variations in his environment. But, as a matter of empirical fact, it is always conceivable that in any selected group the variations of the two factors may be in some degree correlated, possibly owing to some indirect process of causation. Let us therefore begin by taking as our two rectangular axes, first the horizontal line OH, to represent, by its direction and length, the relative influence of heredity (*H*), and a second line perpendicular to it, namely, the vertical axis, OR, to represent the influence of environment (*R*) in so far as it is uncorrelated with heredity (see Fig. 1). Their lengths will accordingly be drawn proportional to the square roots of the corresponding variance, i.e. to the standard deviations.

We now require a supplementary component to represent environment, in so far as it operates in the same direction as heredity. This can be represented by the line RS parallel to OH. The joint effects of these two environmental influences will then be indicated by a line, OS. Evidently the observed test-marks, OT, will be found by completing the parallelogram, OSTH.

Problem. As analysed in Table I, the results obtained yield figures indicating (i) the length of the base, OH, (ii) the length of the diagonal, OT, and (iii) the height of the parallelogram, OR. The point T can therefore be determined by describing a circle with centre at O and radius OT cutting RT in T. Hence our main task will be to determine the lengths of the lines RS and OS (which are needed to measure the influence of the components S and J and can be calculated by Euclid I, 47) and the angles between the various vectors (which are needed to measure the various correlations).

Working Procedure. The suggested steps in the calculations are simple if a little lengthy. To begin with, we find the lengths of the lines OH, OR, and OT by taking the square roots of the variances: e.g. for the crude test measurements (col. 1 of Table I), $OH = \sqrt{77.06} = 8.7784$; $OR = \sqrt{5.91} = 2.4305$; $OT = \sqrt{93.57} = 9.6732$.

Since $\cos ROT = OR/OH = 0.2513$, we can at once determine the angle ROT ($75^\circ 26'$) and its sine (0.9679); and, since $RT = OT \sin ROT$, we can now calculate $RS = RT - ST = RT - OH = 9.3627 - 8.7784 = 0.5843$. The remaining steps are obvious. We compute in turn: $\tan ROS = 0.2403$, angle

ROS = $13^{\circ} 29'$, $\cos \text{ROS} (r_{rs}) = 0.9724$, $\text{SO} = 2.5002$, $\text{angle SOT} = 61^{\circ} 57'$, $\cos \text{SOT} (r_{st}) \times 0.4703$, $\text{angle SOH} = 76^{\circ} 31'$, $\cos \text{SOH} (r_{sh}) = 0.2332$, $\text{angle TOH} = 14^{\circ} 34'$, $\cos \text{TOH} (r_{th}) = 0.9679$. The square of OS gives the variance of the joint environmental effects, viz. $2.5002^2 = 6.2514$; and the last three cosines specify the three correlations which we chiefly desire to know—namely, those between (i) the systematic effects of the environment and the tests, 0.4703, (ii) the systematic effects of the environment and heredity, 0.2332, and (iii) heredity and the tests, 0.9679.

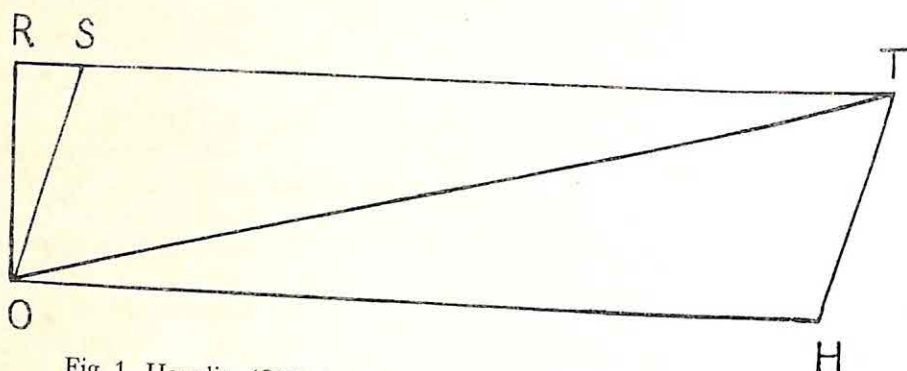


Fig. 1. Heredity (OH), Environment (ST), and their Resultant (OT).

Alternative Analysis of Variance. We can now re-analyse the variance of the reduced total, σ_t^2 , in terms of the chief contributory factors just specified. Writing σ_h^2 for the variance contributed by heredity, σ_j^2 for the variance of the joint environmental effects, and r_{hj} for the correlation between the two, we have, by the familiar formula,

$$\sigma_t^2 = \sigma_h^2 + \sigma_j^2 + 2r_{hj} \sigma_h \sigma_j;$$

that is, for the Crude Test Marks,

$$\begin{aligned} &= (8.7784)^2 + (2.5002)^2 + 2 \times 0.2332 \times 8.7784 \times 2.5002 \\ &= 77.06 + 6.25 + 10.26 = 93.57. \end{aligned}$$

In the same way we obtain for the Adjusted Assessments,

$$\begin{aligned} &(9.3574)^2 + (2.4031)^2 + 2 \times 0.0317 \times 9.3574 \times 2.4021 \\ &= 87.56 + 5.78 + 1.42 = 94.76. \end{aligned}$$

We may tabulate these results as shown in Table II.

TABLE II. ALTERNATIVE ANALYSIS OF VARIANCE

	Crude Test Marks	Adjusted Assessments
Heredity		
Environment	77.06	87.56
Effects of correlation between heredity and environment	6.25	5.78
	10.26	1.42
Total	93.57	94.76

This procedure enables us to subdivide the total variance in such a way as to show not only the proportion contributed (i) by genetic influences and (ii) by environmental influences (both 'systematic' influences, i.e. those operating in the same direction as heredity, and 'random', i.e. those operating in statistical independence), but also (iii) the contribution due to the joint effects arising from the fact that the two main components may be more or less correlated. We observe that, for the crude test measurements, this third contribution is fairly large—over 10 per cent, whereas for the adjusted assessments it is almost negligible, amounting to between 1 and 2 per cent only.

Empirical Verification. From the initial figures for the component variances, it will be noted, we are able by this method to deduce an estimate for the correlation between environmental influences, on the one hand, and the assessments for intelligence, on the other. Taking the crude test marks, we obtain a correlation of 0.470; and on substituting the adjusted assessments we obtain 0.277. It should be possible to check these two theoretical estimates by comparing them with the empirical coefficients. All that we require are objective assessments for the relevant environmental conditions. It was impossible to procure the necessary figures in the case of every individual child in the entire group. But, as noted in our previous paper, we were able to secure trustworthy ratings for the 'socio-economic status' of a large number of households, which together form a fairly representative sample. These include ratings for (a) the material (i.e. financial and hygienic) conditions of each family, and (b) the cultural (i.e. educational and motivational) background.

(a) For the material environment the correlations with the assessments of intelligence are (i) with the crude test marks 0.288 and (ii) with the adjusted assessments 0.226. (b) For the cultural environment the correlations are (i) with the crude test-results, 0.453, and (ii) with the adjusted assessments, 0.315. These later coefficients agree quite well with the theoretical values we have deduced; and they thus provide an added confirmation for the validity of the analysis.

Causal Influences and Path-Coefficients. Finally, we may note that the square roots of the two variances, represented by the lines OH and OS, may be written $\beta_{th.e}$ and $\beta_{te.h}$, and can be regarded as the partial regression coefficients of the actual assessments (i.e. in the diagram, of the test results) on heredity and environment respectively. Sewall Wright terms such quantities 'path-coefficients': these two, it should be observed, are 'one-way' path-coefficients. The correlation between heredity and environment, r_{he} , he would also regard as a 'path coefficient': but it is essentially a 'two-way' coefficient. With this interpretation we can think of heredity as influencing (or as capable of influencing) the test-results along two paths simultaneously: first, by the direct path represented by the regression coefficient $\beta_{th.e}$, and secondly by an indirect path, leading first from heredity to environment (as indicated by the presence of a significant correlation between them), and thence by the path leading from environment to the test-results as indicated by the other regression coefficient, $\beta_{te.h}$. An

analogous interpretation may be adopted, at least as a theoretical possibility, in regard to environment; environment may influence the test-results both directly and indirectly, the indirect influence passing first through the same two-way path as before, but in the opposite direction.

However, when we attempt to examine causal influences in this concrete way, it seems scarcely satisfactory to stop at this point. There are two obvious questions which still remain undecided. Both are questions which statistical analysis *by itself* is utterly incompetent to solve.

In the first place, as we have noted, environment may in some degree operate, or tend to operate, in the same direction as heredity. Suppose we could show that this effect occurred merely among those families where the parents, in virtue of their own inheritable intelligence, had been able to provide a better cultural environment for their children; many would be tempted to argue that, instead of crediting this effect to environment, we ought to regard it as yet another instance in which heredity is the primary cause. Whether such a statement is reasonable would depend on how the initial question itself was phrased; and no doubt, in the sweeping form we have just used, the conjecture would almost certainly be misleading: sometimes parents of no great intelligence do succeed, either by making financial sacrifices or (in times past) by the aid of inherited wealth, in providing exceptional intellectual opportunities for their children; these opportunities would almost inevitably increase the *apparent* intelligence of their children (though perhaps not so much as is commonly believed). This is not the place to discuss such issues in and for themselves. We mention them to show that they are of a type which can only be decided by specific investigations, and that in such investigations statistical analysis will play merely an auxiliary part.

Secondly, as the suggestions of several correspondents indicate, many feel that it would be desirable to split what we have called "the effects of the correlation between heredity and environment", and assign appropriate portions definitely to one or the other of the two correlated variables. In dealing with cases like the present, Heilman, for example, has suggested that the joint contribution should be divided in proportion to their direct contributions¹. This seems rather arbitrary. We ourselves believe that, without further data obtained expressly to settle such an issue, it is impossible to carry the analysis further than we have done; and the requisite data could only be procured by a line of inquiry quite different from that with which we are here concerned. In any case the results recorded in the last of our tables (Table II) seem to give a reasonably adequate answer to the limited questions with which we began; and these complex supplementary issues must be deferred to a later occasion.

¹ Heilman, J. D., 'The Relative Influence upon Educational Achievement of Some Hereditary and Environmental Factors', *XXVIIth Yearbook Nat. Soc. Study Educ.* (1928), pp. 35-65; cf. *id.*, 'Factors determining Achievement and Grade Location', *J. Genet. Psychol.*, XXXVI, 1929, pp. 435-456.

THE THEORY AND USE OF THE SUCCESS-RATIO

By DAVID A. WALKER

Deputy Director of Education, Fife

Abstract. The success-function defined in this paper has been found useful in follow-up studies of school children where entrance marks have been normally distributed. It provides a rapid method of estimating the product-moment correlation between entrance mark and success when the criterion for success is of the pass-or-fail type.

I. INTRODUCTION

No system of selecting pupils for courses of secondary education is, or can be, completely successful while children preserve their right to develop in their own ways and at their own speeds. But in predicting success in the secondary school different batteries of tests show different degrees of efficiency; and much of the research in this field has consisted of follow-up studies undertaken to compare the validities of tests and other measures used at the selection stage. The concept of success-ratio is one that does not appear to have been used to its full extent; and the present paper is an attempt to develop its theory and application.

The assumptions that must be satisfied if a full use is to be made of it are : (i) that the entrance marks of the whole group under consideration are normally distributed; (ii) that success (or its lack) is a normally distributed variable, those pupils above a certain level in that attribute being successful, while those below that level are unsuccessful; (iii) that the correlation between entrance mark and success is linear.

The applicability of the first assumption to any set of data can readily be tested, while the second and third assumptions can be checked by the methods described in this paper. It will be observed that the question of correction for attenuation is not raised. The method thus avoids the difficulties both of suitable assumptions and of additional calculation experienced by previous experimenters (e.g. W. G. Emmett [1], W. McClelland [2]).

II. THEORY OF THE SUCCESS-RATIO

If the assumptions just stated are satisfied, it is possible to calculate the expected value of the success-ratio for any given entrance mark. Let x represent the entrance mark, measured in units of standard deviation from zero mean. Let y represent the success-failure variable, similarly measured. Let $y = k$ be the line dividing success from failure, values of y greater than k representing success. The correlation surface is

$$Z = \frac{1}{2\pi\sqrt{(1-r^2)}} \exp\left\{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)\right\},$$

where r is the correlation between the variables x and y . For a given value of x , the frequency of success is therefore

$$\int_k^\infty Z dy = \frac{1}{2\pi\sqrt{(1-r^2)}} \int_k^\infty \exp\left\{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)\right\} dy,$$

which can be shown to equal

$$\frac{\exp(-\frac{1}{2}x^2)}{2\pi\sqrt{(1-r^2)}} \int_k^\infty \exp\left\{-\frac{1}{2(1-r^2)}(y-rx)^2\right\} dy.$$

Now $\exp(-\frac{1}{2}x^2)/\sqrt{(2\pi)}$ is the total number of candidates with mark x in the group. Hence

$$\frac{\text{Number successful at level } x}{\text{Total number at level } x} = \frac{1}{\sqrt{(2\pi)}\sqrt{(1-r^2)}} \int_k^\infty \exp\left\{-\frac{1}{2(1-r^2)}(y-rx)^2\right\} dy$$

This fraction may be termed the 'success-ratio' at x .

$$\text{Let } \frac{y-rx}{\sqrt{(1-r^2)}} = -u, \quad dy = -\sqrt{(1-r^2)} du.$$

Limits of integration are

$$-u = \frac{k-rx}{\sqrt{(1-r^2)}} \text{ and } \infty, \quad \text{i.e. } u = \frac{rx-k}{\sqrt{(1-r^2)}} \text{ and } -\infty.$$

Hence the success-ratio at level x

$$\begin{aligned} &= \frac{1}{\sqrt{(2\pi)}\sqrt{(1-r^2)}} \int_{\frac{rx-k}{\sqrt{(1-r^2)}}}^{-\infty} \{-\sqrt{(1-r^2)}\} \exp(-\frac{1}{2}u^2) du \\ &= \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\frac{rx-k}{\sqrt{(1-r^2)}}} \exp(-\frac{1}{2}u^2) du. \quad (1) \end{aligned}$$

It is therefore the integral of a normally distributed variable with zero mean and unit variance between the limits $-\infty$ and $(rx-k)/\sqrt{(1-r^2)}$. In other words, it is a function of r , x and k only. Consider this as a function of x . Put

$$u = \frac{ru' - k}{\sqrt{(1-r^2)}}, \quad du = \frac{r}{\sqrt{(1-r^2)}} du'.$$

Limits of u' are now $-\infty$ and x , and success-ratio at level x

$$= \frac{r}{\sqrt{(2\pi)}\sqrt{(1-r^2)}} \int_{-\infty}^x \exp\left\{-\frac{(ru' - k)^2}{2(1-r^2)}\right\} du', \quad (2)$$

which is the integral of a normally distributed variable u' with mean at $(ru' - k)/\sqrt{(1-r^2)} = 0$, i.e. $u' = k/r$, and standard deviation $\sigma' = \sqrt{(1-r^2)}/r$, both measured in σ units from mean of ellipse.

This function with mean at k/r and standard deviation equal to $\sqrt{(1-r^2)}/r$ we shall term the 'success-function'. The theorem established is that the success-ratio at any entrance mark, x , standardized on the whole group is the integral of the success-function from $-\infty$ to x .

III. APPLICATIONS

(a) *The rapid estimation of success-ratios.* This application is most readily illustrated by an example. Suppose that the entrance marks are I.Q.s centred on 100 with standard deviation 15, and that the correlation between I.Q. and success is 0.8; suppose also that the line dividing success from failure is at $k = 1$ S.D., i.e. 16 per cent succeed.

Then the success-function is the probability function with mean at $k/r = 1/0.8 = 1.25$ corresponding to an I.Q. of $100 + 1.25 \times 15$ or 119, and standard deviation of $\sqrt{(1-r^2)}/r = 0.75$ corresponding to an I.Q. range of 0.75×15 , i.e. 11.

From this distribution the theoretical success-ratio at any precise I.Q. level can be estimated. For example, pupils with I.Q. levels 119 have a success ratio of 0.50; those with I.Q. 130, i.e. 11 points above this mean, have a success-ratio of 0.84; those with I.Q. 108 have a success-ratio of 0.16, and so on. These values may be read from tables of the normal distribution, or from a straight line drawn on probability paper.

TABLE I. SUCCESS-RATIOS OF 1949 GROUP AT UNIVERSITY ENTRANCE LEVEL

Entrance Mark	Number in Group	Number gaining University Entrance	Success-Ratio
100 and over	8	5	0.625
95-99	29	12	0.414
90-94	105	27	0.257
85-89	225	32	0.142
80-84	377	10	0.026
75-79	491	3	0.006
70-74	514	2	0.004
Below 70	2581	0	—
Total	4330	91	—

Mean Mark 65.3. Standard Deviation 14.3.

(b) *Estimation of correlation from success-ratios.* The usual situation is that the success-ratios are known, but not the correlation between entrance mark and success; and the experimenter wishes to estimate the correlation from the data. The discussion will be clearer if an example is given from actual practice. In 1949, 4,330 pupils in a Scottish county were transferred to courses of secondary education, their courses having been partly determined by the entrance mark in a battery of tests and other measures; five years later, 91 had obtained a Scottish Leaving Certificate of a sufficiently high standard to admit them to a Scottish University. The performance of the whole group is shown in Table I. To save space the lower part of the table has been telescoped; the distribution of entrance marks was in fact roughly, though not completely, normal.

Method A. From the theory given above it is obvious that an estimate of the correlation coefficient r can be obtained from each success-ratio separately,

provided that k is known. For example, the second group which has a mean entrance mark of 97 has a success-ratio of 0.414. The upper limit of u (eqn. (1)) is -0.217 ; hence

$$(rx - k)/\sqrt{(1 - r^2)} = -0.217. \quad (3)$$

Now $x = (97 - 65.3)/14.3$ in standardized units; and k is the line cutting off 91/4330 from the tail of the normal curve, i.e. $k = 2.034$. (Inspection of Table I suggests that no suitable candidate has been excluded from the successful group by the operation of the pass mark. This point will be considered in greater detail in a later section.) The solution of eqn. (3) (taking the positive value of the root since it represents a standard deviation) is $r = 0.87$. Estimates of r could be obtained in the same way from the other entrance levels, and these could be combined. The most accurate way of doing this is shown under Method C below.

Method B. The most rapid method of estimating the correlation coefficient from a set of success-ratios is by the use of the relation between r and the standard deviation of the success function. If the success-ratios are plotted on probability paper against the mid-points of the entrance-mark intervals, the points should lie on a straight line if the assumptions underlying the method are satisfied. From this straight line, the entrance mark for which the success-ratio is 0.50 and that for which the success-ratio is 0.16 are read off. These give the mean of the success-function and the point one σ unit below the mean. The difference between these measures is therefore the standard deviation of the success function. When it is standardized by division by the standard deviation of the total group of entrance marks, it equals $\sqrt{(1 - r^2)}/r$; and r is rapidly found.

In practice the points are not absolutely collinear; and a line of best fit may be drawn by eye. When this is done with the data of Table I, it is found that a success-ratio of 0.50 occurs at entrance mark 98.0 and one of 0.16 at entrance mark 89.5. Hence

$$\sqrt{(1 - r^2)}/r = (98.0 - 89.5)/14.3, \quad \text{i.e. } r = 0.86.$$

Alternatively, we may convert the success-ratios into the corresponding normal deviates; and the graph may then be drawn on ordinary graph paper. If the assumptions are satisfied, the graph is once more a straight line, and the standard deviation of the success-function corresponds to unit change in the normal deviate.

Method C. For a more exact method of estimating the correlation coefficient from a set of success-ratios we must fit the best line to the series of points, giving each set of observations the appropriate weight. This is ensured by the method devised by Bliss (cf. Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Introduction to Tables IX and XI). Here we are concerned with the success-ratios of pupils with different entrance marks. Bliss's method not only enables us to calculate the slope of the line relating the normal deviate to entrance mark, and so to calculate the correlation underlying the success-ratios, but also supplies a value for chi-squared indicating whether the assumptions of normality and of linear correlation are in agreement with the data.

Details of the calculations are given in the Appendix (pp. 46f below). When the method was applied to the data of Table I, the correlation coefficient obtained was 0.85. The value of chi-squared was 5.8, giving a probability of at least 0.5 that the deviations could be attributed to sampling errors.

IV. THE EFFECT OF THE PASS-MARK

If entrance to the course is limited to those with entrance marks $x > h$, then the successful candidates will be those in the correlation ellipse for whom $y > k$. The theory outlined in section II of this paper applies only to those for whom $x > h$, and modifications must be made in the methods of estimating r . Method A is no longer applicable, as k is not known. Methods B and C can still be used for all groups above the admission mark.

Let us apply Method C to the data of Table II which shows the number of pupils from the 1949 group who obtained a certificate of any type after six years of secondary education. The admission mark to the course was 75, but some pupils with lower marks were admitted on appeal and others transferred to the Leaving Certificate course after having done well in other courses. A dotted line is drawn across the table to separate these candidates from the others.

TABLE II. SUCCESS-RATIOS OF 1949 GROUP IN MINIMUM CERTIFICATE AFTER 6 YEARS

Entrance Mark	Total Number in Group	Number Successful	Success-Ratio
100 and over	8	8	1.000
95-99	29	19	0.655
90-94	105	56	0.533
85-89	225	102	0.453
80-84	377	91	0.241
75-79	491	56	0.114
.....			
70-74	514	10	0.020
65-69	543	2	0.004
60-64	568	1	0.002
Below 60	1470	0	0.000

When Method C is applied to the figures above the dotted line, the value of r obtained is 0.79 and the value of chi-squared is about 5.5, giving a probability of 0.25 that the deviations are within the bounds of sampling errors. Let us now suppose that no candidate below the pass mark of 75 had been admitted to the course. The numbers successful in the class intervals 70-74 and below would then have been zero in each case. The application of Method C would furnish a correlation coefficient of 0.86; but the value of chi-squared would become 27.7, and the consequent probability about 0.001. The deviations would be too great to be attributed to sampling errors and the assumptions, e.g. linearity of correlation, would be no longer tenable. No importance could be attached to the correlation so obtained.

It is interesting, though rather irrelevant to the main purpose of this article, to note that when the actual numbers of successes are used, the value of r obtained from the whole table is 0.84, and the value of chi-squared is 12.7; which corresponds to a probability of 0.20. The flexibility of the selection scheme has therefore restored the position. The correlation obtained from the whole table is greater than that obtained from the part above the line, but it is not possible to state whether the difference is significant, as the standard errors of correlation coefficients obtained by these methods have not been worked out.

V. SUMMARY

1. From a set of success-ratios there can be derived a success-function which can be used to predict other success-ratios and also to estimate the correlation between entrance mark and success.
2. Methods of estimating the correlation between entrance mark and success are given and illustrated by examples.
3. The advantages of the use of the success-function in estimating correlations are :
 - (i) the assumptions are few, simple, and easily tested ;
 - (ii) the intricacies of correcting correlation coefficients for attenuation are avoided.

APPENDIX. *Details of Calculations for Method C*

For ease of calculation the mid-points of the class intervals of the entrance marks may be denoted by the variable x , chosen so that $x = 0$ at entrance mark 87. For each success-ratio we find the appropriate probit, i.e. the normal deviate increased by 5 to avoid negative values. These probits may be obtained from a table of the normal distribution; a convenient table is provided in Fisher and Yates' Table IX. They may be described as the experimental probits, and are shown in Column 5 of Table III. We next plot the experimental probits against the entrance mark x and draw the best fitting line by eye. From this line we obtain provisional probits for use in the next part of the calculation. The provisional probits chosen should be simple values, occurring in Fisher and Yates' Table XI. We now calculate from the data given in Table XI the working probit y . This equals the minimum working probit plus the success-ratio times the range. For example, the working probit for $x = 3$ is $3.7186 + 0.625 \times 2.5573$.

Table XI also gives the weighting coefficient to be used with each provisional probit. The actual weight w is obtained by multiplying the weighting coefficient by the total number of pupils at that level.

We now calculate $S(w)$, $S(wx)$, $S(wy)$, $S(wxy)$, $S(wx^2)$, and $S(wy^2)$, and from them obtain

$$\begin{aligned}\bar{x} &= S(wx)/S(w) = -0.3408; \\ \bar{y} &= S(wy)/S(w) = 3.593; \\ A &= Sw(x - \bar{x})^2 = S(wx^2) - \bar{x}S(wx) = 537.549; \\ B &= Swy(x - \bar{x}) = S(wxy) - \bar{x}S(wy) = 305.708; \\ C &= Sw(y - \bar{y})^2 = S(wy^2) - \bar{y}S(wy) = 179.684.\end{aligned}$$

TABLE III. DATA FOR DETERMINING b AND χ^2

Entrance mark	x	Total Number	Number Successful	Experimental Probit	Provisional Probit	Working Probit (y)	Weight (zw)
100-104	3	8	5	5.3	5.2	5.3169	5.016
95-99	2	29	12	4.8	4.7	4.7831	17.864
90-94	1	105	27	4.3	4.2	4.3563	52.815
85-89	0	225	32	3.9	3.7	3.9650	75.600
80-84	-1	377	10	3.1	3.2	3.0809	67.860
75-79	-2	491	3	2.5	2.7	2.5371	37.120
70-74	-3	514	2	2.3	2.2	2.3698	12.644
65-69	-4	543	0	—	1.7	1.4194	3.334
60-64	-5	568	0	—	1.2	0.9522	0.670

The slope of the line relating probit to entrance mark is $b = B/A = 0.5687$. Unit change in probit therefore corresponds to a change of $1/0.5687$ in entrance mark. This must be standardized in terms of the class interval of 5 that has been used and of the standard deviation of the whole group of entrance marks, i.e. 14.3. The corresponding change of entrance marks is therefore $5/(0.5687 \times 14.3)$ or 0.6148 which thus represents the standard deviation of the success function. This, according to the theory, equals $\sqrt{(1-r^2)}/r$, from which $r = 0.85$.

Also $\chi^2 = C - bB = 5.83$. The appropriate number of degrees of freedom is 7, since two constants have been fitted, and the corresponding value of P is between 0.5 and 0.6.

REFERENCES

- [1] EMMETT, W. G. (1945). *An Inquiry into the Prediction of Secondary School Success*. London: University of London Press.
- [2] McCLELLAND, W. (1942). *Selection for Secondary Education*. London: University of London Press.

JOURNAL COMMITTEE

As instructed by the Council of the British Psychological Society, the Journal Committee, at its meetings in 1956, carefully re-examined the questions of title, typography, and price; and now wishes to thank the numerous contributors and readers of the journal who helped by communicating their views. In its final recommendations the Committee strongly supported the new arrangements for printing, and urged that there should be no change either in the title or in the price of the journal. It suggested that the arguments of those critics who found the articles too technical might be partly met if contributors could be persuaded to submit papers on a wider range of topics, e.g. experimental, elementary or expository articles, and notes or abstracts from theses.

Council deferred a full consideration of the Committee's recommendations until its meeting on July 6, 1957, i.e. until after the first issue of the current year had been published. The decisions then reached were reported by the Editor to the Committee in his review for the first half year of 1957 (to July 31), which the Committee recommended should be published.

The report stated that the changes in typography, which the Editor had proposed in 1954 (11-pt Imprint with 1-pt leading, and abolition of small type) had now been sanctioned. Largely as a result, the previous deficit had been converted into a profit; and accordingly, after some discussion, Council had agreed that the price should at present not be raised.

Council, however, feared that subscribers might now feel that they were getting an inadequate return for their subscriptions, since increase in type-size and spacing, following on the temporary reduction to two issues a year, would mean that less material could be presented in an issue of 64 pages. The Editor explained that it was now proposed to return to the plan of issuing three parts a year, and that a benefactor had guaranteed a fund to cover any loss that could not be met out of the profits. Part III would include an Index for the first ten volumes. The alternative would be to increase the number of pages in each part. The former proposal was accepted by Council. But the Editor would still welcome the views of readers.

Since the Publications Committee had recommended that editors should serve for a limited period of years only, Council had nominated Dr. Whitfield as his successor.

It was agreed that a formal expression of thanks should be recorded to the American benefactor for her aid, to contributors and readers for their continued help and support, and to all those (including the new printers and publishers) who had enabled the journal to reach a position where no serious deficit need any longer be feared.

Correspondents are asked to note that the Officers of the Society are no longer the Publishers of the Journal. Communications regarding material for publication should be sent to the Editor; questions relating to business matters should be addressed to the Managing Editor.

CHARLOTTE BANKS,
Secretary.

CRITICAL NOTICE

Secondary School Selection : A British Psychological Society Inquiry.

Edited by P. E. VERNON. London: Methuen & Co., Ltd., 1957. Pp. 216. 15s.

This survey will be widely welcomed for its summary of recent literature on the important issues raised: and the conclusions put forward may no doubt be taken as representing the general views of British psychologists on one of the most heated controversies of the day¹. The book deserves the special attention of the statistical psychologist, not only because it shows how serviceable his techniques have proved, but still more because it sets forth a number of specific problems on which his help will be essential.

The report is the joint production of a Working Party which consisted of about a dozen members of the British Psychological Society. With few exceptions each contributed one or more chapters; and, as was almost inevitable, the treatment varies appreciably in clarity and outlook. Thanks no doubt to the care taken by the Editor and Secretary (Professor Vernon and Miss Doris Lee), the statistical issues are clearly brought out and the statistical evidence given full weight.

The Preface warns the reader that "the memorandum cannot represent the views of every member of the Society; and no attempt is made to conceal that on some points differences of opinion or emphasis have arisen". In certain cases a non-committal compromise appears to have been accepted. In others the contradictions are left unresolved; and it is frankly admitted that "divergences of outlook remain". As several critics have already pointed out, it is perhaps to be regretted that the Working Party included none of the many experienced school teachers who are members of the Society nor any educational official or administrator whose memory could reach back to the practical difficulties which hampered the early reformers and which are so largely responsible for the shape the problem takes today. One consequence is that the issues involved are discussed as if they were primarily psychological, whereas it is the administrative requirements that must determine any workable solution².

¹ In view of the intense interest which the problem has recently aroused and of the volume of correspondence received by the editors since the Society's inquiry was published, the Journal Committee suggested that the book should be reviewed at some length. The task has been entrusted to two writers who are fully acquainted with modern statistical methods and have also had first-hand experience of the practical issues either as teacher or as administrator. The Editor regrets that space does not permit the publication of the letters in this issue: but the reviewers have been asked to take note of the points so raised. The full report recently published on the same subject by the National Foundation for Educational Research was received too late for review at the same time; but it will be discussed in the following number. The reviewers would welcome further correspondence, which will be published or summarized at the same time.

² Miss Lee, who has been good enough to read this review in manuscript, points out that the terms of reference received by the Working Party referred primarily to 'psychological aspects of secondary school selection' (Editor).

The Inquiry starts with a discussion of the 'Background of Selection', and then passes to a 'History and Survey of Present Procedures'. These two chapters (as one of our correspondents observes) "have apparently been drafted by writers who are more familiar, it would seem, with the recent findings of American psychologists than with the earlier efforts of English investigators". No reference is made to the work of such leading educationists as Ballard, Kimmins, Sir Robert Blair, Sir Percy Nunn, and Sir Michael Sadler, whose endeavours to place 'secondary school selection' on a scientific basis were far more influential than the few who are mentioned by name. The bibliography of nearly 150 numbers provides a useful conspectus of recent publications. But only ten authors are cited who made contributions to the problem before 1930; and of these, six are American.

The writer of the first chapter tells us that it was "the Butler Education Act that gave us the blue print for the future of our schools which won unusually unanimous approbation"; and he apparently supposes that the proposals then made represented "an attempt to catch up with the changes in the structure of society following the war". In point of fact, there was hardly anything in the Act which had not been suggested twenty years before; and much of the present difficulty is due to the fact that even among progressive educationists the attitude towards those suggestions has been far from unanimous. It would be truer to say that the 'blue print' was originally provided by the most celebrated of all the Reports of the Consultative Committee of the Board of Education—the report on *The Education of the Adolescent* (1926). Yet, strangely enough, the list of references at the end of the book omits all mention of it.

The inspiration behind that report came largely from Sir Percy Nunn, Principal of the Institute of Education, and by far the most eminent educationist of his day. It was he who so effectively urged the substitution of statistical studies and statistical data for the impressionistic arguments that had previously prevailed; it was his papers and memoranda which had argued so persuasively for the "clean cut across our educational system at the age of 11+ as in accordance with the results of recent psychological research"; and it was he and his colleagues at the Institute who pressed for the use of psychological tests as an aid to selection and secured the abolition of the 'Rule 3(b)' of the Board's *Regulations for Secondary Schools*, which had for so long hindered their adoption. For many years he had urged¹ that the term 'secondary' should be extended to cover all types of 'post-primary education', and that a third type of secondary school, in addition to the 'technical' and 'grammar', should be introduced, namely, "schools with a realistic or practical trend to be known as 'Modern Schools'" (a name suggested as equivalent to the German term *Realschulen*). And it was his suggestion that, in order to be sure of catching in the net the ablest members of each group, the proportion actually transferred to the grammar schools should be raised to between 15 and 20 per cent—the actual figure varying according to the population of the area. At the same time both Nunn and the Consultative

¹ T. P. Nunn, *Education: its Data and First Principles* (1920), pp. 237f.

Committee envisaged many other types of school or class according to the requirements and traditions of each district.

We cannot therefore help feeling that the Working Party has represented the problem as much newer and far simpler than it really is. Moreover, in all its reports the Consultative Committee treated its various proposals as involving a process of *allocating* pupils to suitable schools rather than of *selecting* them for favoured types of school. Unfortunately, as the title of their 'Inquiry' implies, the Working Party seem rather to have retained the popular notion of the procedure, namely, that it is a kind of competitive selection for grammar schools. "At least half the child population," they say, "is competing for grammar schools which will only accommodate one-fifth." But that, plausible though it may seem to many, suggests a decidedly misleading picture of the fundamental problem. The real need at the moment is not so much more grammar schools for those who at present cannot get in, but more schools of a technical, commercial, or other specialized types. And, although at the end of the historical survey there is a short 'Note on Selection for Technical and Commercial Schools' reminding the reader that "this topic must not be omitted altogether", this aspect of the matter, and the numerous psychological questions which it involves, are in point of fact never really faced.

Having outlined the general problem and described in somewhat vague terms 'the Diversity' and 'the Adaptability of English Education', the writer of the opening chapter goes on to ask why after all parents are so anxious to see their boys and girls accepted by a grammar school. His answer is that, when their child fails to win acceptance, parents from the middle classes "feel that they are losing face", and as a result often prefer a relatively inefficient private school to the "stigma of going to a modern school". Teachers who have enjoyed a wider first-hand knowledge of the reasons that actuate most parents apparently think this explanation a little unfair. Thus several of our correspondents are of the opinion that (to quote one writer) "snobbery as a motive is far more frequent among the manual classes than among parents of the professional or middle classes: how often does one hear a father say: 'I only let him go there to show that he's as good as any other kid; I never meant him to stay'". As Mr. Skeet has said, owing perhaps to some lurking sense of inferiority, the working class parent "does not so much want a *suitable* education for his child, but only that his child shall go to as good a school as their neighbour's, i.e., to the one that is most esteemed"¹. On the other hand, middle class parents (if we may trust a survey recently carried out by Mr. R. T. Graham) tend as a rule to adopt a far less egoistic attitude: with them the commonest objection to the modern school is, "not that it entails a social disgrace, but rather a fear lest the child may pick up rude manners, a low vocabulary, a slovenly accent, slangy and ungrammatical speech, and above all disgusting words and ideas". By going to a grammar school with a long-standing tradition, the boy will (so the parent hopes) not only escape such risks, but "acquire something of the cultural,

¹ D. V. Skeet, *The Child of Eleven* (University of London Press, 1957).

moral, and social ideals, the spirit of loyalty and sense of responsibility and public service that are part of the very atmosphere of the English public schools and old-established grammar schools ”.

In their history of the psychological study of the problem the writers are evidently more at home. Yet even here they seem to underestimate the work done in this direction by the more progressive administrative authorities themselves—notably the inquiries and reports undertaken by the London County Council and by the Board of Education itself. It is to these bodies and to certain outstanding members of their staffs—Young, the Secretary of the Consultative Committee, Daniel, the Chief Examiner to the London County Council, Kimmins, its Chief Inspector, and Nunn, the Principal of its Training College (later known as the Institute of Education)—far more than to the academic writers whose names are better known, that the real initiative was chiefly due.

As the Preface points out, the contribution of the psychologist as such starts with “the pioneering work of Galton, who was interested in discovering innate abilities of a high order as well as in assessing cases of subnormality”. But the systematic use of intelligence tests as part of the educational routine began with the London County Council—the first education authority to appoint a psychologist as an Officer on its educational staff. The duties of the psychologist were not merely to apply psychological methods, but to investigate their practical value and, where necessary, develop and recommend new procedures.

For the assessment of subnormal children, the most obvious procedure was to use ‘individual’ tests, such as those devised by Galton and Binet. For the scholarship examinations, intended to select brighter children for central schools and grammar schools, the application of individual tests (the only type available in the early days) would have been far too expensive; and accordingly a scheme of written tests for the majority, checked by teachers’ ratings, with oral examinations for borderline cases, was put forward for trial. The idea of employing ‘group’ tests started with the experiments on ‘higher mental processes’ carried out by Burt in 1909; and he was apparently responsible for most of the stock types still used in examinations for secondary schools and described in the Working Party’s report—Opposites, Analogies, Classification, Reasoning, Number Series, short Completion tests, and the like.¹ In view of recent complaints of coaching and ‘test-sophistication’, it is surprising that so few fresh types have been invented during the past forty years. Here, we feel, the Working Party might well have suggested new types of problem, or at least emphasized the need for devising tests of a less antiquated kind.

In these earlier years most academic psychologists looked on such proposals with marked disfavour. Spearman urged the substitution of “simple laboratory tests, as befits the rigour of scientific work” (e.g. tests of sensory discrimination individually applied); Thomson questioned the whole notion of a general factor of intelligence. And the prolonged controversies over the very existence

¹ Cf. *Experimental Tests of Higher Mental Processes and their Relation to Intelligence*, *J. Exp. Pedag.*, I, 1911, pp. 93–112.

of 'general ability' and of 'special aptitudes' served only to augment the natural scepticism of officials and committees¹. Nevertheless, the experiments carried out by the London County Council aroused widespread interest among administrators; and, as soon as the first world war was over, the group tests standardized by Burt for London were tried by education authorities in Blackpool, Leicester, Northumberland, the West Riding of Yorkshire, and elsewhere, as well as for entrance examinations at Cheltenham Grammar School and at Rugby. All who used them pronounced favourably on the results. But the first authorities to introduce a separate intelligence test as part of their regular procedure were those whose areas included a mixture of both rural and urban schools: as one of the reports explained, "in country districts many pupils may be quite as well endowed in natural ability as those in the towns; but, for various reasons, notably the low standards often prevailing in the village schools, they are apt to reach a comparatively poor level in ordinary examinations of a scholastic type".

In 1920 the Consultative Committee took up the problem at the request of the Board. As regards the theoretical issue the Committee unhesitatingly accepted the distinction between 'inborn ability' and 'acquired attainments', and the double assumption of an 'innate general ability' and 'innate special aptitudes maturing rather late'; and concluded that "a number of children qualified to profit by secondary education are at present excluded because, for one reason or another, their attainments are not equal to their abilities Examinations for entrance to secondary schools should be designed primarily to discover ability rather than to test attainments". Accordingly "tests of 'intelligence'" were recommended for adoption, in addition to other methods, on the ground that they provided "a species of written or oral examination designed to test ability rather than attainments".

It is evident, however, that the Committee were not entirely satisfied with the use of the word 'intelligence' to describe the tests used for this purpose. The reasons are perhaps most clearly set forth in the memorandum by Burt referred to in their report. A distinction is there drawn between "the various purposes for which so called intelligence tests are used: in examining the borderline defective it is desirable, before certifying the child, to ascertain, so far as practicable, whether his backwardness is the irremediable consequence of an innate lack of general ability or merely the incidental result of remediable causes; but in examining a pupil for admission to a secondary school, it is less important to assess his innate ability as such, since that is but one of many factors determining his capacity to profit by the education there given. For instance, the verbal and scholastic bias of many tests, which should be wholly eliminated in tests for the defective, may actually prove an advantage in selecting pupils for a comparatively bookish and academic type of education It is, therefore, unfortunate that the phrase 'intelligence test' is used to describe two different types of test having two different functions. The teacher in the elementary school, when receiving new pupils at the age of 7 or 8, rightly seeks a method of

¹ See Board's report on *Psychological Tests of Educable Capacity* (1924), especially appendix IX, which was largely concerned with these underlying issues. However, by the time the report was printed, both Thomson and Spearman had begun to modify their earlier views.

testing innate ability, since at that stage what he wants to know are the inborn potentialities of each child; but that is no longer the primary purpose of scholarship examinations at the age of eleven or later." Here "the real criterion is, not, how accurately does this procedure enable us to assess inborn potentiality, but how accurately does it enable us to predict the child's capacity to benefit by the type of education proposed". The Working Party state that the situation now obtaining in Britain "clearly calls for some estimate of the child's potentialities at the end of the primary stage". We should prefer to see this done at the beginning. And for the test used to estimate the child's suitability for the grammar or modern school we should prefer some such name as 'general classification test'.

The equivocality of the term 'intelligence' is responsible for much confusion. As the Working Party go on to remind us, "intelligence suggests something inherited; thus the parent is much more emotionally involved in the child's I.Q. than in his performance at attainment tests or ordinary examinations". Accordingly one of their chief aims has apparently been to allay these parental fears, which are the source of so much ill-informed opposition. They seem to have thought that the best way to do so would be to minimize the importance of hereditary differences and the possibility of determining them. But after all it is the existence of these innate differences that causes the problem and makes it of such national importance; and surely one can deny the need to test a thing for practical purposes without denying its existence on theoretical grounds.

The chapter which deals with the actual use of so called 'intelligence tests' is much more lucidly written than the rest. The author frankly accepts the results of factor analysis, but rejects the extreme and antagonistic views of both Spearman and Thurstone. "Though factor analysis does not enable us to pin down the nature of intelligence," he says, "it has nevertheless developed to a stage where it can effectively analyse and classify the component abilities The factor *g* (for *general* ability) is," we are told, "an empirical fact: as Burt pointed out, it is the Highest Common Factor that can be extracted from analysing any set of test scores"; but at the same time "he provided evidence for numerous group factors or subtypes of ability." In spite of Spearman's arguments therefore we must recognize both.

The writer then goes on to inquire "how far do intelligence tests measure inborn ability?" This, however, as we have already implied, is really a side-issue; and in any case the answer must surely be that all depends on which test we have in mind and on the way it is applied. On the general problem several extensive investigations have recently been published in the Society's own journals (cf., this *Journal*, IX, p. 126 and refs.) to which the Working Party does not refer. It would appear that with a good group test of the ordinary type about 77 per cent of the variance is attributable to genetic factors; and with supplementary checks even this figure can be raised. For evidence on this issue, however, the Inquiry relies mainly on earlier American research. In the end the views of "left-wing writers" who hold that "intelligence test results largely or

mainly reflect social class differences" are firmly repudiated; and Lawrence's studies of children transferred to orphanages are quoted as "proving beyond doubt that such social class differences [as the tests reveal] are in some measure genuinely innate".

So far, therefore, the conclusions seem to agree both with those of the Consultative Committee and of the ordinary teacher¹. But, as we read on, we encounter those unreconciled "divergences in outlook" that we are warned to expect. The Preface, as we have seen, urges the need for estimating the child's innate potentialities, and describes how education authorities are now using "tests designed to give an indication of innate ability as distinct from scholastic performance". Yet, on a later page we are assured that such tests "do not measure innate ability". Again in another passage, drafted presumably by another hand, we are told that "by intelligence we mean the more general qualities of comprehending, reasoning, and judging, which have been *picked up* without much specific instruction"; and it is therefore argued that, if I.Q. tests give a fair indication of capacity for advanced education, that is "not because the capacity is inborn, but because it has been more or less stabilized in the previous eleven years". It is in our view perfectly true to say that what we want to test in such examinations is 'stabilized capacity for further education'; but to accept this view does not mean that we must first deny genetic differences or the possibility of assessing them.

The mention of 'I.Q. tests' leads to a discussion of the time-honoured question of the constancy of the I.Q. "From about 6 to 11", it is said, "the I.Q. from a good individual test is unlikely to alter by more than about 7 points on the average"; (that should reassure teachers who wish to determine at an early age which children are innately handicapped and which have the potentialities of good scholars). Hence "one can definitely conclude that the I.Q. is sufficiently stable to make possible useful predictions in the majority of cases over the period of primary or secondary schooling". But once again, having thus confirmed the general view, the writers then proceed to make such large concessions to the rival doctrine that the ordinary reader is bound to feel bewildered. In certain cases, it is said, the I.Q. "may change 30 or 40 points; . . . many children with I.Q.s in the 50 to 70 range may show themselves capable later of coping with education in the ordinary school". Let us apply this to children tested at 10: it means that a pupil who then had a mental age of 8 might at 12 have a mental age of over 14, and that "many" who had mental ages between 5 and 7 could later be satisfactorily educated in the ordinary school. Surely in such cases the more reasonable inference would be that the I.Q.s were obtained either by an incompetent tester or with unsuitable tests.

¹ A few years ago Professor Vernon submitted a questionnaire to a number of teachers, asking (among other things) how they understood the term 'intelligence': the definition 'inborn all round mental efficiency' (or its equivalent) was accepted by 84 per cent and rejected by only 9 per cent, the rest declining to commit themselves: (*The Bearings of Recent Advances in Psychology on Educational Problems*, 1955).

But all this is really irrelevant. Whether or not group tests of the usual type are of practical help in the selection examination is an issue to be decided, not by inquiring how far "the I.Q., properly measured, is stable or not", but by ascertaining how far the tests actually do predict the children's subsequent performances. And to the statistical reader the portion of greatest interest will be the chapter dealing with "validity"—a problem which for some reason is discussed before the description of the tests and other procedures whose validity is to be assessed. For assessing it the writer advocates the use of multivariate analysis and the discriminant function; but he does not explain how these methods would be applied or quote their results. The substance of his review seems excellent; but his style is by no means lucid. If we interpret him correctly, he holds that two main problems call for special study. First, how far does the present system "select or reject the right pupils"—or, as he puts it later, "to what extent do the predicting tests place pupils in the right order of their potentiality to succeed?" Secondly, what should be "the size of the grammar school group"—or, in other words, "does the system take too many or too few?"

But there is, as he points out, a prior question to be settled: how are we to judge the success? The verdicts of the secondary school teacher are, he believes, not sufficiently trustworthy. Nor is he satisfied with an external examination like the 'G.C.E.'. The best criterion, he thinks, would be an internal examination conducted by the school. Personally we are tempted to argue that profiting by what is intended to be a liberal type of education should be judged, not by any sort of examination, but on a genuinely liberal basis. However, the evidence actually available rests, not on a single criterion, but on "the cumulative effect of numerous investigations": (the more important results are tabulated in a useful appendix); and the outcome "greatly increases our confidence in the efficiency of the selection procedure". But at this point the Working Party seems to become unexpectedly optimistic. "Throughout this report," says an inserted footnote, "we have assumed that the correlation of an efficiently conducted selection procedure with a perfectly reliable scholastic criterion is close to 0.90." This refers to "validities for over 5 years". And the practical conclusion drawn is that "little gain in predictive efficiency can be expected from further tinkering"¹.

But is not the figure suggested a little too high? Since mental growth during adolescence is itself liable to unpredictable fluctuations and since a child's eventual achievements must be influenced by these and many other unforeseen accidents, there can be no such thing as a 'perfectly reliable scholastic criterion' in any literal sense. Moreover, even the reliability coefficient of the selection procedure seldom rises as high as 0.90. The validity coefficients which approach this figure have all been obtained by 'boosting' the observed value to

¹ In the final chapter on 'Recommendations' the essential conclusion is more carefully expressed: "the intelligence test is so consistently successful that it should not be dispensed with, though it might be improved".

allow for selection. The observed correlations, we are told, are "typically of the order of 0.45". Consequently a theoretical value of 0.90 would virtually double the observed coefficient.

However, having assumed a validity of 0.90 and a selection-ratio of 20 per cent, the writer infers that the number of pupils wrongly placed would be about 10 per cent of the total; one quarter of those selected (5 per cent of the total) would not really deserve their places, while another 5 per cent (1 in 16 of the failures) would presumably have made good in the grammar school had they been given the chance. Now this method of assessing the results implies a further assumption which seems scarcely tenable. Why must we suppose that the number of pupils who genuinely deserve places (as judged by the criterion) must be exactly equal to the number who are given the chance (as judged by the selection procedure)?

Before we can properly settle this point, we must take up the second of the problems mentioned above; and this in turn involves two distinct questions.

First of all, how many should we *actually* transfer? The Consultative Committee, it may be remembered, suggested an entry of about 20 per cent, with wide variations for different areas. The Working Party as a whole favours a much higher proportion, namely, 25 to 30 per cent. On the other hand, the only educational administrator included among their members suggests a lower figure—15 per cent. The main principle governing the Working Party's argument is apparently an overriding anxiety lest injustice should be done to the less favoured individuals. But in our efforts to give the benefit of the doubt to every questionable case, we may do serious injustice to the community: for we shall almost inevitably lower the standard of teaching and culture in the schools that receive these poorer specimens: and we shall be frittering away our limited financial resources on pupils who at best are rather dubious speculations¹.

But there is another question which we ought to ask: how many are there in the total population who *really can* profit by this more costly type of education? When a teaching institution in a University selects candidates for training with a view to a degree in medicine, let us say, it does not expect every one of those accepted to pass the final examination: the teaching course itself provides a period of probation. And the number who eventually pass will be much smaller than the number who were originally accepted. Similarly, in accepting candidates for a grammar school, there are two distinct numbers to be borne in mind: (i) we want, so far as we can, to ensure that the ablest members in the total population (the top 2, 5, or 10 per cent or whatever it may be) shall receive an education suited to their ability and aptitudes; but (ii) in order to do so with as few omissions as possible, it will be necessary to admit as many as we can

¹ As many as 18 per cent now leave the grammar school before 16, and 49 per cent at 16; and the proportion of early leavers is, as a rule, much greater among pupils from the manual classes and in areas which select a high proportion for entrance: cf. Central Advisory Council for Education, *Early Leaving*, H.M. Stationery Office, 1954. The social and cultural dangers involved in the recent changes in educational organization are ably discussed in *The Uses of Literacy* (1956) by Richard Hoggart, himself a member of the working classes, who profited in this way.

manage, having due regard to the need (a) not to waste money on those who are bad bets, (b) not to lower the standards of the school, and (c) not to impose undue strain on the less able pupils who are accepted. It follows that the total number of transfers should not be equal to, but appreciably larger than, the total number of successes as estimated for the entire population.

The disagreement over the number proposed for transfer by different members of the Working Party seems to reflect a basic weakness in the whole approach. Throughout there is a tendency to assess the practical value of a selection procedure (or of a component test) solely by reference to its theoretical validity. Now what an education committee wants to know is whether the merits of any new procedure will justify the expenditure that it will presumably entail. And this question can hardly be answered unless we can first attempt some fairly definite estimate for both costs and merits.

But, it may be asked, how can this be done except in terms of rough impressionistic arguments? A careful scrutiny of the literature reveals plenty of parallel examples and a variety of available devices. However, the point is so important, and yet so completely overlooked in the Report, that it seems desirable to cite a concrete illustration. In submitting his budget for expenditure on psychological testing or the like, the Psychologist to the London County Council used regularly to draw up a kind of balance sheet for expected profit and loss, and in doing so assigned to each item a weight depending on its own probability. For instance, in the memorandum on 'The Use of Intelligence Tests and Rating Scales in Examinations for Free Places', Burt begins with the well known formula for determining 'mathematical expectation' (given in every schoolbook on higher algebra), viz., ' $x_i = p_i G - q_i L$ ', where i denotes any one of the alternative procedures to be considered, G the gain achieved by selecting a suitable candidate, L the loss entailed from selecting an unsuitable candidate, and p_i and q_i their respective probabilities with this procedure. These expressions are then expanded in terms of their chief components and their respective probabilities, e.g. the cost of the procedure used the amount the Council is willing to spend on educating a suitable candidate at a grammar school, the borderline proposed for a pass, and so on. In summing the 'expectations' for the several candidates, allowance must be made for the fact that the distinction between pass and fail is relative rather than absolute; and this is effected in the usual way by weighting each candidate according to his marks and weighting each mark by the regression on the criterion. Where two or more alternative policies are to be compared, Burt takes typical combinations of the component variables, calculates the total expectation for each, chooses the minimum, and maximizes that¹.

¹ The use of discrete combinations involves interpolation, which seems a little circuitous. The general principle, it is explained, is based on that used by Edgeworth for his 'hedonic calculus' (maximizing the minimum values of the expected excess of pleasure over pain: F. Y. Edgeworth, *Mathematical Psychics*, 1910). The method followed would seem in some ways to be the reverse of that which has since figured largely in so called 'acceptance procedures', viz., minimizing the maximum 'risk'. However, in the case of the hedonic calculus (we are told) "minimizing 'risk' could most easily be achieved by doing nothing at all—a policy too often adopted in actual practice".

The conclusions reached afford an excellent instance of the fallacy of basing practical recommendations on theoretical validities alone. The procedure with the highest validity was one which included a separate intelligence test, as well as the usual papers in English and Arithmetic and the teachers' marks; yet it is shown that, in a well-organized educational area like London, the *additional* information supplied by the intelligence test was too slight to warrant the additional cost; while in mixed areas (partly rural and partly urban), it appeared, such a test might amply justify its cost.

In Britain, where the funds at their disposal have been strictly limited, the pioneers of applied psychology, whether in education or in industry, have always paid special attention to questions of profit and cost. And today, even in America, arguments of this kind are now a commonplace. As one educationist has put it, "in every investigation where practical issues are involved, statistical theory should insist that considerations of cost and of the consequences of the decisions should be taken into account"¹. It is therefore surprising that this aspect of their problem is not touched upon in the Working Party's 'Inquiry'.

A short chapter is devoted to the assessment of attainments in school subjects and to the supplementary techniques available for borderline cases, such as the study of children's interests, the assessment of character qualities, and the value of interviews and oral examinations. An 'oral interview' (with individual tests) was advocated by the Consultative Committee as a useful adjunct to the ordinary examination; and a recent survey shows that approximately 70 per cent of the various County Councils now adopt some kind of interview for doubtful cases. This the Working Party vigorously condemns. But none of their reasons seems very convincing.

First, we are told, an interview "puts a strain on many children"; and case-studies are quoted to illustrate how a dread of the examination may induce nervous stress. But, as most school doctors would agree, nervous strain is far commoner among the dull but conscientious pupils *after* they have been transferred to the grammar school than it is among candidates for the examination. A tactful interview would do much to obviate such consequences. Secondly, "interviewers are liable to be impressed by the wrong things, such as deportment or accent". But it surely should be part of the trained interviewer's technique not only to put the child at his ease, but also to guard against his own personal prejudices. Thirdly, "wartime research confirmed psychologists' doubts as to the consistency of the interviewers and the validity of their judgments". However, unlike teachers, most wartime psychologists were new to such tasks; and the attempt to judge the personality of adults is a problem quite different from assessing the ability and character of children of eleven.

Finally, it is said, "whenever conclusions derived from interviewing children are allowed to outweigh evidence from test-scores and teachers' estimates, the accuracy of prediction is likely to be lowered". But the object of the interview

¹ M. A. Girshick (*Rev. Educational Research*, XXIV, 1954, pp. 448f) who gives a useful survey of modern statistical procedures of this type.

is not to secure evidence which shall "outweigh" other estimates, but rather to check and supplement them, particularly where the results of the tests and the estimates of the teacher are ambiguous or at variance. At the same time, it affords a convenient occasion for collecting and coordinating further information about the child's aptitudes, interests, behaviour and past history, for hearing the views of the parents, and above all for making sure that no accident of ill health or transitory disturbance was responsible for the low marks obtained for the written examination. Every grammar school master would prefer to see the borderline candidates himself before any decision is taken as to whether they are suitable for his school, while the head of the primary school would, if present, feel reassured that no case was turned down without full consideration. Certainly the N.U.T. Committee of 1949 looked on the old fashioned interview with disfavour. But their objections scarcely apply to what is sometimes called the 'new style of interview'¹. In short, the disadvantages alleged seem to be reasons, not for discarding the interview, but for improving it.

The report concludes with thirty-two italicized recommendations. The majority are excellent in substance, though not always felicitously worded. At present, of the local education authorities in England and Wales, barely one quarter enlist the services of a psychologist. "Psychologists," says the report, "should uphold their claim to be able to diagnose the most suitable form of schooling." But the ability to 'diagnose' a suitable school is primarily an educational rather than a psychological problem; and the majority of educational psychologists are more familiar with the diagnosis of dull, backward, and neurotic pupils than with the 'diagnosis' of schools. What the psychologist can do, provided he possesses the requisite training and experience, is to help in the construction, standardization, and application of tests, and offer advice on individual children whose ability or character seems problematic. But much of this advice could and should be given long before the stage of selection is reached.

With the more practical recommendations this is scarcely the place to deal. But the detailed suggestions as to the theoretical problems calling urgently for research may well be brought to the notice of our readers. The following seem the most important: "the broader aspects of benefiting from a grammar-school course, and their predictability," "the development of intellectual capacities in adolescence", "'old-type' and other forms of examination or assessment which, without loss in predictive efficiency or undue unreliability in scoring, might have beneficial rather than harmful backwash effects," and finally "the general scaling of school marks and estimates": (on this last problem there is a most instructive statistical note in one of the appendices). In all these inquiries, the Working Party rightly "insists that the procedures should be planned and supervised by persons with adequate psychological and statistical training".

To the statistical psychologist perhaps the most interesting feature of the whole inquiry will be its incidental and unanimous vindication of statistical

¹ See Mr. Skeet's account of the 'new style interview' (as worked out, e.g. by the Northumberland authority in collaboration with Professor Thomson): *The Child of Eleven*, pp. 131-135.

procedures as the most objective and profitable means available for studying complex practical issues like that of selection. At the same time both its merits and its shortcomings illustrate the supreme importance of combining the theoretical approach of the statistical psychologist with the practical experience of the teacher and administrator. Many of the investigations cited have been carried out by research-students, unfamiliar with the conditions in the classroom and the home, and often working under purely academic guidance. It therefore seems high time that a larger number of education committees should follow the example already set by one or two of the most progressive, and appoint experienced psychologists as members of their own administrative staffs with full authority to undertake systematic surveys and cooperative studies, scientifically planned, of the various problems arising within their areas.

A. G. HOUGHTON

D. R. MORRIS

BOOK REVIEWS

Statistical Methods and Scientific Inference. By R. A. FISHER. London: Oliver and Boyd, 1956. Pp. viii + 175. 16s.

This is a book which every statistical psychologist should read. The author explains that in his two previous volumes he had in view two practical aims: in *Statistical Methods for Research Workers* his object was to modify traditional statistical procedures (based on the theory of infinitely large samples) in order to meet the needs of experimental work; in *The Design of Experiments* the aim was to improve the quality of the data obtained in such work by considering in advance the logical nature of the procedures involved. In these earlier volumes, however, space had not permitted a full discussion of "those refinements of the reasoning process which have come clearly into view with a study of these two aspects of scientific work". It is with these further refinements that Sir Ronald Fisher's new book is mainly concerned.

Karl Pearson's statistical and biological concepts are somewhat severely criticized. His mathematics are described as clumsy and sometimes misleading (e.g. in problems of distribution); and "in his dispute with Bates on over Mendelian inheritance", we are told, "he was the bull to the skilful matador". The newer theories of decision functions are recognized as useful in the fields of industrial and commercial psychology for which they were originally designed; but their adoption as aids to forming correct scientific conclusions is firmly rejected.

W. R. NORTON

An Introduction to Cybernetics. By W. ROSS ASHBY. London: Chapman and Hall, Ltd., 1956. Pp. x+295. 36s.

Dr. Ashby's aim has been to provide an introductory textbook on 'cybernetics'. His definition of the term is "the science of control and communication in the animal and the machine: in a word, 'steersmanship'". As he explains, the present volume is complementary to his earlier work, *Design for Brain*. It falls into three main parts: the first discusses processes developing within a single system; the second, processes of communication between one system and another; and the last, the problem of regulation and control in biological systems. No knowledge of mathematics is required beyond elementary algebra; and each section is followed by graded exercises with hints and answers.

Starting from a few elementary concepts—'operands' (i.e. states or properties of 'systems'), 'operators', 'transitions' (i.e. the changes effected by the operations), repeated transitions, and the like—the author shows how each can be made more precise by formal definition and classification. A complex set of transitions, produced by operating simultaneously on a number of 'operands' within the same system, is called a 'transformation', and is represented symbolically by the multiplication of matrices. The series of states exhibited by such a system at successive times is expressed by the successive application of the appropriate transformation, and is said to define a 'trajectory' or 'line of behaviour'.

In applying this formal procedure to the solution of practical enquiries, the first step is to relate the behaviour of the concrete physical mechanism to the abstract properties of the symbolic expressions—that is, of the matrices and the matrix-operations. We are then introduced to the conception of a 'determinate machine'. This is defined as 'that which behaves in the same way as a closed single-valued transformation'. The definition, it will be noted, turns not on the nature of a material structure, but on its mode of behaviour. The point is illustrated by what the electrical engineer would call 'the problem of the black box': we are confronted with an apparatus possessing terminals for input to which he may apply any stimulation he pleases, and terminals for output, which will enable him to observe the results: the box itself cannot be opened; and the problem is to describe its behaviour or infer its inner structure in the light of the observations so made.

In Part II we are introduced to a mechanism whose behaviour is stochastic, i.e. "determinable only in a statistical sense"; and the concepts and principles explained in the first part of the book are progressively developed so as to form a basis for the theory of such processes as coding, communicating information, regulating behaviour, feed-back in its various forms, and the attainment of stability or equilibrium. Throughout this type of analysis the more familiar questions of energy play little or no part; and often the most important variables take the form of probabilities. Moreover, unlike the methods generally adopted in mathematical physics, those proposed by Dr. Ashby are not restricted to metrical characteristics or to processes that are continuous and linear. Since in the biological sciences the behaviour investigated is frequently non-linear and non-continuous, and at times not even metrical, such techniques are naturally of special interest to workers in those fields.

Psychology itself is a biological subject; and a number of psychological problems are briefly discussed by Dr. Ashby to illustrate his points. A special characteristic of the 'mechanisms' studied by the psychologist is that they seek goals. And goal-seeking is, Dr. Ashby tells us, "an example of behaviour that is stable around a state of equilibrium"; nevertheless, as he points out, not all the examples of stability so attained are 'goals', since not all of them are 'desirable'. Here the reference to 'desire' apparently implies the notion of 'value' or 'satisfaction' which, it would seem, can never be reduced to purely mechanistic concepts. Generally, he suggests, "it should be possible for the cybernetically minded psychologist to devise a topological approach which will enable him to get the information he needs". The book ends with an interesting section on "amplifying the intelligence" (in a rather special sense of that term). Problem-solving, he maintains, is essentially a matter of appropriate selection; and, since power of selection can be amplified, "it seems to follow that intellectual power, like physical power, can be amplified". As an instance of such amplification he cites the way "gene patterns grow up to form a brain".

Dr. Ashby's psychological illustrations have been somewhat severely criticized by psychologists themselves; and this has led many to suppose that his methods would be of little use in psychology. The present reviewer would dissent strongly from that verdict. The proposals outlined in his book might perhaps have carried more conviction had he referred in fuller detail to researches already carried out by psychologists, where similar concepts and procedures have already been employed with success. Stout, for examples, over half a century ago asked "whether we can find a mode of physical self-determination, which, like the apperceptive and conative processes studied by the psychologist, is directed towards an end, and ceases when the end is attained"; and found it in the process of 'recovering equilibrium after disturbance' by means of 'adaptation to irregularly varying outer condition' through 'indirect' (or 're-entrant') 'self-determination' (i.e. through feedback). The temporal sequence of states which results he terms a 'vital series' (an inappropriate phase borrowed from Avenarius) (*Analytic Psychology*, I, pp. 151f.).

Lewin attempted a 'topological' approach (*Principles of Topological Psychology*, 1936); but at that date the theory of analysis situs was not sufficiently advanced for much progress to be made. Burt's treatment of the subject, familiar to most readers of this *Journal*, comes nearest to that envisaged by Ashby. In his terminology Ashby's 'systems' become 'persons', the 'operands' 'traits', the 'temporal series' 'occasions', and the 'operators' 'factors' (i.e. matrix multipliers containing 'weights' or 'loadings'). If necessary, qualitative traits may be studied instead of metrical by substituting association coefficients for correlation coefficients. Following Stout, Burt regards the essential process to be the recovery of equilibrium after disturbance; but instead of a single stimulus or response we have patterns, expressed by vectors or matrices. The use of matrix models, taken over by Burt from Sheppard, is of long standing in psychology. The application of the 'principal axis transformation' then enables the investigator to express the whole process in 'canonical' form; and the use of O-technique, based on time, in place of the more familiar T- or P-techniques renders it possible to "deal with dynamic instead of merely static problems". Thus the processes of learning and growth are expressed by repeated multiplication by a factor matrix consisting of 'transitional probabilities', leading, in the simplest case, to a Markov chain (also suggested by Ashby); and what Ashby calls 'trajectories' are expressed by 'autocatalytic curves'¹. These detailed similarities between the procedures worked out by two independent investigators seem to offer strong confirmation of their value.

One common criticism advanced by psychologists is that the problems attacked along these lines by statistical investigators (e.g. those of mental inheritance, learning, and the like) cannot be solved until the underlying conditions have been fully elucidated by the experimental geneticist or the neurologist. To this objection the analogy of the engineer studying his 'black box' seems to afford a convincing reply: after all one can discover a great deal about the behaviour of a new wireless set by twiddling knobs and noting the results without ever peering inside the case of looking at the diagram of the wiring. The psychological problems themselves are of such practical importance that it is urgently desirable to reach at least a provisional solution without waiting half a century for the internal mechanisms to be revealed in detail. For those who are contemplating research in these fields Dr. Ashby's book forms one of the most suggestive contributions published during recent years.

T. F. RAMSAY

¹ Cf. this *Journal*, II, pp. 95f., III, pp. 166f., IV, pp. 179f., and IX, pp. 49f. For some of the analogies between factorial methods and cybernetic methods I am indebted to a lecture given by Professor Burt at Worcester.

INDEX TO VOLUME X

INDEX OF AUTHORS

	PAGE
Bernyer, G. : Psychological Factors : their Number, Nature, and Identification	17
Burt, C., and Howard, M. : Heredity and Intelligence : A Reply to Criticisms	33
Burt, C., and Howard, M. : The Relative Influence of Heredity and Environment on Assessments of Intelligence	99
Dunsdon, M. I., and Roberts, J. A. F. : A Study of the Performance of 2,000 Children of Four Vocabulary Tests	1
Hotelling, H. : The Relations of the Newer Multivariate Statistical Methods to Factor Analysis	69
Howard, M., and Burt, C. : Heredity and Intelligence : A Reply to Criticisms	33
Howard, M., and Burt, C. : The Relative Influence of Heredity and Environment on Assessments of Intelligence	99
Lewis, W. B. : Professor Lewis M. Terman	65
Roberts, J. A. F., and Dunsdon, M. I. : A Study of the Performance of 2,000 Children on Four Vocabulary Tests	1
Stuart, A. : The Comparison of Frequencies in Matched Samples	29
Walker, D. A. : The Theory and Use of the Success-Ratio	105
Wrigley, C. F. : The Distinction between Common and Specific Variance in Factor Theory	81

INDEX OF SUBJECTS

Common and Specific Variance in Factor Theory, The Distinction between : Wrigley, C. F.	81
Frequencies in Matched Samples, The Comparison of : Stuart, A.	29
Heredity and Intelligence : A Reply to Criticisms : Burt, C., and Howard, M.	33
Heredity and Environment on Assessments of Intelligence, The Relative Influence of : Burt, C., and Howard, M.	99
New Multivariate Statistical Methods to Factor Analysis, The Relations of the : Hotelling, H.	69
Psychological Factors : their Number, Nature, and Identification : Bernyer, G.	17
Success-Ratio, The Theory and Use of the : Walker, D. A.	105
Terman, Professor Lewis M. : Lewis, W. B.	65
Vocabulary Tests, A Study of the Performance of 2,000 Children on Four : Dunsdon, M. I., and Roberts, J. A. F.	1

BOOKS REVIEWED

Ashby, W. Ross : <i>An Introduction to Cybernetics</i>	126
Cronbach, L. J., and Gleser, G. C. : <i>Psychological Tests and Personnel Decisions</i>	64
Fisher, R. A. : <i>Statistical Methods and Scientific Inference</i>	125
Sholl, D. A. : <i>The Organization of the Cerebral Cortex</i>	64
Vernon, P. E. (ed.) : <i>Secondary School Selection</i>	113

"GRIFFIN'S STATISTICAL MONOGRAPHS & COURSES"

Edited by M. G. KENDALL, Sc.D.

A new series of publications planned to meet the need for a form of book, at moderate cost, which will make accessible, to a group of readers, specialized studies in statistics or special courses on particular statistical topics.

- | | | |
|--------|---|--------------|
| No. 1. | ANALYSIS OF MULTIPLE TIME-SERIES
M. H. QUENOUILLE | 24s. net |
| No. 2. | A COURSE IN MULTIVARIATE ANALYSIS
M. G. KENDALL | 22s. net |
| No. 3. | THE FUNDAMENTALS OF STATISTICAL REASONING
M. H. QUENOUILLE | In the press |

THE MATHEMATICAL THEORY OF EPIDEMICS

By NORMAN T. J. BAILEY, M.A.

Reader in Biometry, University of Oxford.

A valuable survey, with the main emphasis on the biometrical and epidemiological rather than the mathematical aspects.

General introduction—Historical sketch of mathematical epidemiology—Epidemiological principles—Deterministic theory—Stochastic theory: continuous-infection models—Stochastic theory: chain-binomial models—The measurement of latent and infectious periods—Recurrent epidemics and endemicity—The detection of infectiousness—Summary and conclusions—Index.
36s. net

CHARLES GRIFFIN & COMPANY LIMITED
42 DRURY LANE LONDON WC2

OCCUPATIONAL PSYCHOLOGY

Editor: ALEC RODGER

October, 1957

Volume 31, No. 4

Engineering Psychology: A British Psychological Society Symposium:

- Introductory Remarks
- Information Theory in the Understanding of Skills
- The Objective Study of Judgment and Decision-Taking
- Discussion of the Preceding Papers
- The Design and Layout of Machinery for Industrial Operatives
- Transfer of Training in Engineering Skills
- Environmental Stress and Its Effects on Performance
- Discussion of the Later Papers

FRANK A. GELDARD
HARRY KAY
C. H. BAKER
DONALD WALLIS
W. T. SINGLETON and R. SIMISTER
W. DOUGLAS SEYMOUR
D. C. FRASER
B. UNGERSON
JOHN CROWTHER

Absence and Turnover in the Divisions of One Company, 1950-1955

The Influence of Group Size and Stability upon the Effectiveness of an Incentive Payment System

An Outline of the History and Work of the Medical Research Council's Industrial Psychology Research Group.

By R. MARRIOTT.

Book Reviews

Other Books Received

Index

Annual Subscription 30 shillings

National Institute of Industrial Psychology, 14 Welbeck Street, London, W.1

BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY

Vol. X Part II

November 1957

CONTENTS

LEWIS, W. B.	Professor Lewis M. Terman	<i>Page</i> 65
HOTELLING, H.	The Relations of the Newer Multivariate Statistical Methods to Factor Analysis	69
WRIGLEY, C. F.	The Distinction between Common and Specific Variance in Factor Theory	81
BURT, C., AND HOWARD, M.	The Relative Influence of Heredity and Environment on Assessments of Intelligence	99
WALKER, D. A.	The Theory and Use of the Success-Ratio	105
JOURNAL COMMITTEE		112
HOUGHTON, A. G., AND MORRIS, D. R.	'Secondary School Selection' (Critical Notice)	113
BOOK REVIEWS		126
INDEX		128

JOURNAL COMMITTEE

CYRIL BURT *Editor*

CHARLOTTE BANKS *and* ALAN STUART *Assistant Editors*

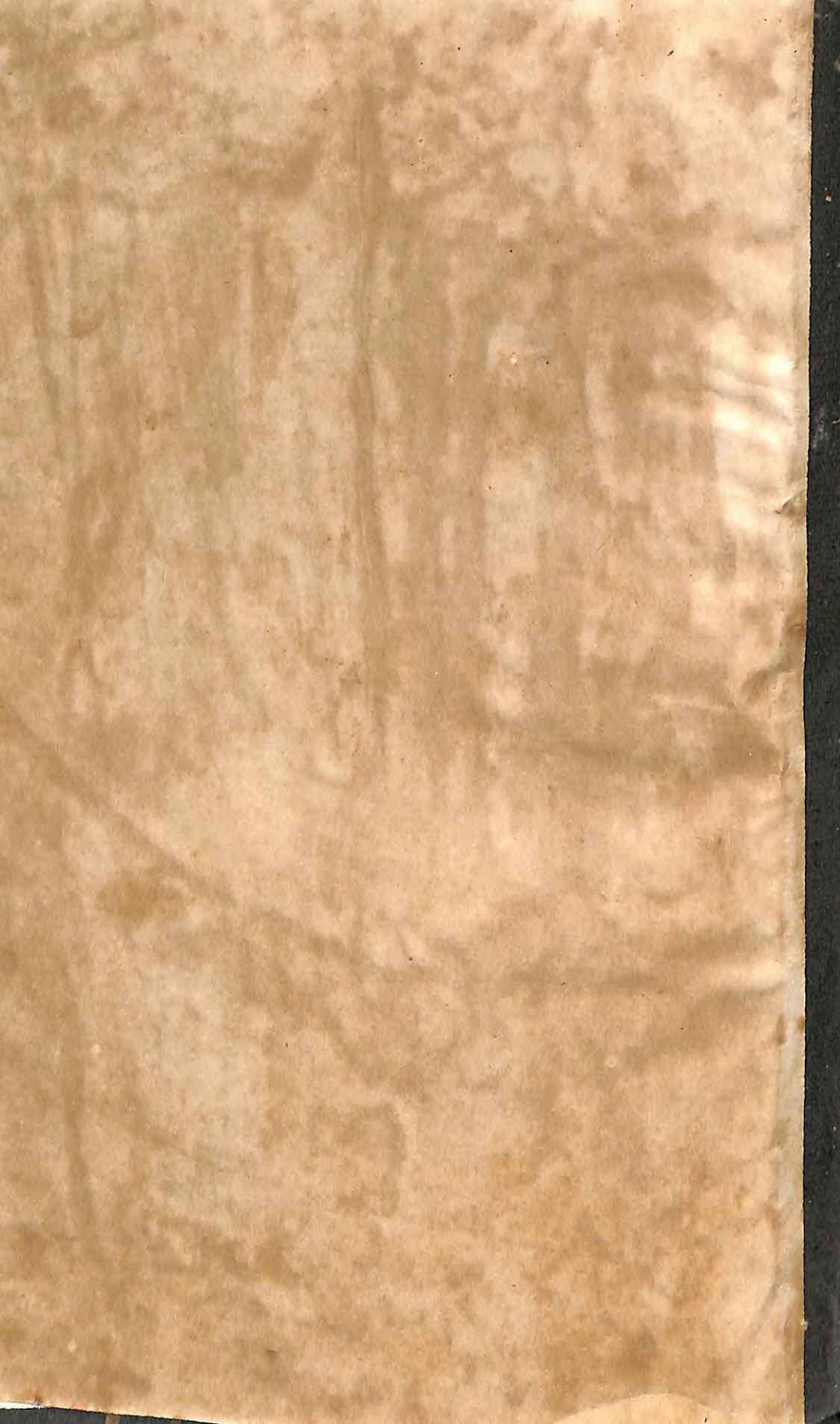
J. W. WHITFIELD *Managing Sub-editor*

and

B. M. FOSS
M. HAMILTON
A. HERON

J. S. SMALL
J. SUTHERLAND
F. W. WARBURTON





MADE BY THE DISABLED
REHABILITATION - INDIA
CALCUTTA-700088